Universidad Autónoma de Madrid
Escuela Politécnica Superior

Grado en Ingeniería de Informática

Trabajo Fin de Grado

VALIDACIÓN Y MEDIDA DE RENDIMIENTO DE MODELOS BAYESIANOS PARA INTERPRETACIÓN FORENSE

Autor: Marta Sánchez Abad
Tutor: Daniel Ramos Castro

Julio 2018

VALIDACIÓN Y MEDIDA DE RENDIMIENTO DE MODELOS BAYESIANOS PARA INTERPRETACIÓN FORENSE

Autor: Marta Sánchez Abad
Tutor: Daniel Ramos Castro

Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio 2018

Resumen

Resumen

En este trabajo se propone una técnica para medir la robustez de un modelo probabilístico de interpretación forense.

El proceso forense consiste en la comparación de muestras de origen desconocido con otras de origen conocido con el fin de ayudar a un juez a tomar una decisión en un juicio tras obtener el valor incriminatorio de dicha comparación.

Para poder ofrecer un apoyo eficaz, el perito forense se basará en las evidencias y utilizará modelos para obtener ratios de verosimilitudes. Sin embargo, para no caer en grandes errores, puesto que no tenemos todo el conjunto de datos posible, solo una muestra de la población, es necesario estudiar el rendimiento de estos modelos y lo robustos que son al variar la cantidad y calidad de los datos.

Para la realización de este proyecto se ha utilizado una base de datos de 62 vidrios con 11346 comparaciones entre ellos, de los cuales solo 3782 serán entre muestras de igual fuente y el resto entre muestras de diferentes fuentes. A partir de esta base de datos y de medidas que muestran el rendimiento, como las curvas de entropía cruzada empírica (ECE), se propone el uso de intervalos de confianza para medir la robustez.

Para obtener esos intervalos se han propuesto dos técnicas diferentes: Bootstrap y Subset Bootstrap, con una diferencia fundamental: mientras que la primera solo se puede utilizar con datos independientes entre sí, en el Subset Bootsrap, se pueden tener grupos de datos dependientes unos de otros, de manera que, dentro de cada grupo habrá datos dependientes entre sí e independientes con el resto de datos de otros grupos. Así, variando la cantidad de datos de las muestras que se utilizan en estas técnicas se obtendrán los intervalos de confianza.

En concreto, con la base de datos que se ha utilizado en nuestros experimentos, puesto que se trata de datos que son dependientes entre sí, la técnica ideal sería el Subset Bootstrap. Con Bootstrap se obtendrán unos intervalos menores pero, se estarían subestimando estos márgenes, puesto que se ha considerado que todos los datos son independientes cuando no lo son. Con Subset Bootstrap, se supondrá esa dependencia entre datos, dando lugar a unos intervalos mayores y más fiables. Esto es debido a que los datos dependientes contienen menos información intrínseca sobre la distribución, y por lo tanto cualquier estadística medida en este conjunto de datos tendrá mayor incertidumbre.

Palabras Clave

Entropía cruzada, modelo bayesiano, validación, discriminación, calibración, curvas ECE, LR, score, evidencia, algoritmo PAV, Bootstrap, Subset Bootstrap, C_{llr}, intervalos de confianza

Abstract

This paper proposes a technique for measuring the robustness of a forensic interpretation probabilistic model.

The forensic process consists of comparing samples of unknown origin with others of known origin to obtain their incriminating power, in order to help judges make a decision.

To provide an effective support, the forensic expert will rely on evidence and use models to obtain likelihood ratios. However, in order to avoid major errors, since we don't have the whole data, only a sample of the population, it is necessary to study the performance of these models and how robust they are when the quantity and quality of the data changes.

For this project we have used a database of 62 glasses of any kind, ranging from bottles to windows, with 11346 glass comparisons of which only 3782 will be between glasses of the same source and the others between glasses of different sources. Based on this database and some measures that show performance, such as the curves of empirical cross entropy (ECE), confidence intervals have been studied to measure robustness.

Two different techniques have been proposed to get those margins of error: Bootstrap and Subset Bootstrap, with one fundamental difference: while the former can only be used with mutually independent samples, Subset Bootstrap allows to form groups, in which the samples are only dependent of others within the group. Thus, by varying the amount of data samples that are used in these techniques we obtain confidence intervals.

With the database that we used in our experiments, since it has dependent data, the ideal technique would be the Subset Bootstrap. Using Bootstrap we would achieve smaller intervals, but we would be underestimating these, since we assume all samples are independent while this is not the case. Subset Bootstrap would achieve more reliable intervals, and since the samples yield less information due to their dependant nature, our intervals will have a higher uncertainty, thus leading to bigger intervals.

Key words

Cross-entropy, bayesian, validation, discrimination, calibration, ECE plots, LR, score, evidence, PAV algorithm, Bootstrap, Subset Bootstrap, C_{llr}, confidence intervals

Agradecimientos

A mi tutor por su paciencia y dedicación.

A mis profesores por su implicación y esfuerzo.

A mis padres y hermano por su confianza y apoyo incondicional.

A mis compañeros y amigos que me han acompañado y ayudado durante toda la carrera.

A mi grupo Yatekomo y a mi mejor amiga y pareja de trabajos, por todas esas horas juntos, las risas y los llantos que nos han acompañado durante toda esta etapa.

Índice general

Índice de figuras

1

Introducción

1.1. Motivación

La ciencia forense se define como el uso de procedimientos científicos para ayudar a tomar decisiones en juicios. En muchos de los casos están incluidas personas a las que se las está juzgando por un delito. Es el juez el que tiene que combinar todas las circunstancias del delito o información a priori con la evidencia analizada por el científico forense y tomar una decisión.

Un típico caso forense es en el que hay una marca incriminatoria de origen desconocido. De esta manera, el papel del forense es dar el peso de la prueba, consistente en los materiales incriminatorios, y los materiales del sospechoso. En la ciencia forense moderna, una forma de cuantificar el valor de una comparación entre dos materiales de un caso forense es utilizar "scores"que sirven para dar una representación cuantitativa de lo que se está midiendo, en muchos casos similitud y rareza en una población. Un ejemplo sería en el que hay una ventana rota en una escena del crimen y además, existe un sospechoso con trozos de vidrios en su chaqueta. El objetivo del perito forense que se encargue del caso sería comparar los vidrios de la chaqueta del sospechoso, con los vidrios de la ventana rota. En ocasiones, tras la comparación se obtendrá un valor llamado "score"de manera que cuánto más alto fuera, mayor sería la similitud de esos dos vidrios.

En la ciencia forense la interpretación de evidencias está basada en los ratios de similitudes o también llamados "LRs"para apoyar la toma de decisiones. Éstos se obtienen, en muchos casos, a partir de la transformación de los scores. Desde principios del siglo XX, la interpretación de dicha comparación se ha llevado a cabo mediante procedimientos simplistas que no tenían en cuenta la incertidumbre presente. Sin embargo, debido a graves errores, estos modelos han evolucionado a modelos más complejos.

Sin embargo, no solo es importante el desarrollo del método, también es importante su validación. Muchos modelos pueden fallar estrepitosamente cuando se les ponen condiciones adversas. Es por eso, que en el presente TFG se estudiarán y propondrán métodos para medir la robustez que utilicen datos cuantitativos para evaluar la incertidumbre de las técnicas empíricas utilizadas.

1.2. Objetivos

El objetivo principal de este trabajo es el de estudiar y probar distintas técnicas que midan la robustez de un modelo que genere LRs. Se probarán en un entorno con escasez de datos. De hecho, es muy típico este escenario de pocos datos en un caso real forense.

Sin embargo, antes de probar estas técnicas será necesario el estudio del proceso forense, en concreto del análisis de evidencias que será en el que nos vamos a centrar. Tendremos que entender el marco bayesiano de decisión, la necesidad de utilizar un LR como medida del valor de la prueba, además de algunas medidas de rendimiento para evaluar si un conjunto de LRs se comporta como debería. Solo entonces, será cuando entenderemos la necesidad de que nuestro sistema sea robusto, y buscaremos las técnicas para comprobar si lo es o no.

El proceso seguido para obtener estas medidas de robustez será el siguiente:

1. En primer lugar se buscará una medida cuantitativa de la robustez de un modelo. La medida con la que trabajaremos serán los intervalos de confianza sobre el rendimiento basado en entropía cruzada empírica (ECE).
2. En segundo lugar, se usarán técnicas para sacar estos intervalos de confianza variando el conjunto de datos, en concreto, la cantidad de muestras de nuestra base de datos. Las técnicas estudiadas y desarrolladas serán el Bootstrap y el Subset Bootstrap.
3. Finalmente, sacaremos conclusiones sobre las anteriores técnicas, con sus respectivas ventajas y desventajas.

1.3. Estructura del documento

La memoria se ha organizado tal y como se describe a continuación:

1. **Introducción**: se describe el problema y la motivación que ha llevado al estudio de este TFG, así como los objetivos definidos para la resolución del problema.
2. **Estado del arte**: se introduce el proceso de comparación forense y sus niveles de abstracción; se explica el marco bayesiano de decisión en el que se basa este trabajo; se hace una introducción a la validación, siendo éste el objetivo principal del proyecto; y, se explican conceptos sobre rendimiento importantes para entender el TFG.
3. **Diseño**: se estudian y explican los conceptos importantes que necesitaremos saber para entender los experimentos: la entropía, las curvas ECE como medida de rendimiento y la importancia de la robustez en un modelo.
4. **Desarrollo**: se introduce y se explica el desarrollo de las técnicas propuestas para medir la robustez.
5. **Experimentos realizados y resultados**: se describen los procedimientos y experimentos llevados a cabo para el estudio de estas técnicas. Se probarán las técnicas de Bootstrap y Subset Bootstrap con nuestra base de datos. En primer lugar se harán pruebas básicas para un mejor entendimiento de estas técnicas, y después se explicarán unas pruebas más complejas.
6. **Conclusiones y Trabajo futuro**: se muestran las conclusiones extraídas a partir de los resultados, así como las líneas de trabajo futuro relacionadas con la investigación de este proyecto.

2

Estado del arte

2.1. Proceso de comparación forense

En cualquier caso forense, tendremos dos tipos de material: por una parte, el llamado **"dubitado"** del que no se conoce si tiene relación o no con el crimen, y por lo tanto, de carácter incriminatorio (p.e., fragmentos de vidrio de un sospechoso dado, huellas dactilares, manchas de sangre, etc.). Por otra parte, tendremos el material **"indubitado"**, que es el material de origen conocido (p.e.: un fragmento de vidrio de una ventana, una huella dactilar registrada, una muestra de sangre, etc.). [1]

Para efectuar el análisis de evidencias tomamos como entradas, los materiales dubitados e indubitados explicados anteriormente, las bases de datos donde se encuentran las características de las marcas indubitadas y la información "a priori" de la escena del crimen. Con todo ello, propondremos una metodología con distintos niveles de abstracción. En la figura 2.1 podemos ver estos niveles y la salida de cada uno de ellos.

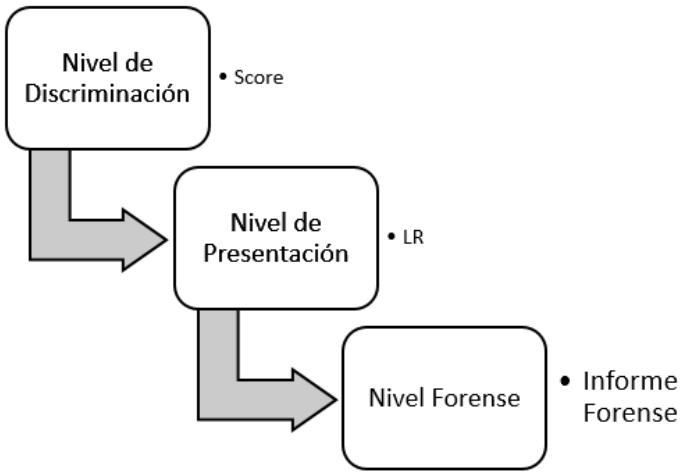

Figura 2.1: *Esquema de los niveles de abstracción para el análisis de evidencias en un caso forense. Adaptado de [1]*

2.1.1. Nivel de Discriminación

El objetivo de este nivel es obtener una puntuación o "score" a partir de la comparación entre materiales dubitados e indubitados. Este score se calculará a partir de la extracción de características de las dos marcas, de manera que, cuanto mayor sea la similitud entre ambas o mayor sea la rareza de dicha similitud, mayor será también esta puntuación. En este nivel es importante una buena discriminación entre dos hipótesis: si las marcas son de la misma fuente o de fuentes distintas.

2.1.2. Nivel de Presentación

El objetivo de este nivel es transformar los scores obtenidos en el nivel anterior en una razón de verosimilitudes (LR). Estos valores son los que representan el peso de la evidencia en un caso forense.

Para llevar a cabo este paso, se utilizan modelos que permiten describir dichas puntuaciones. Aunque se usan muchas técnicas estadísticas y de reconocimiento de patrones, en este trabajo no nos vamos a centrar en ellas, la principal aportación es medir el LR, independientemente del modelo origen.

2.1.3. Nivel Forense

El objetivo de este último nivel consiste en la elaboración del informe forense para el juez con la inclusión del análisis de evidencia junto a pruebas experimentales y explicativas.

2.2. Marco de decisión bayesiano

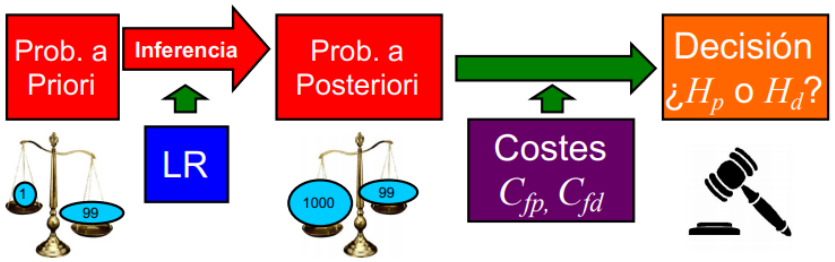

Figura 2.2: *Esquema de la inferencia bayesiana en el marco forense. [2]*

En la figura 2.2 tenemos un esquema representativo del marco de decisión bayesiano. En este esquema vemos como a partir de unas probabilidades a priori y un LR se saca una probabilidad a posteriori, con la cuál a partir de unos costes se obtendrá una decisión. De todo este esquema, el perito forense solo se encarga del LR, es decir, de dar una razón de verosimilitudes a partir de la cuál el juez utilizará junto con las probabilidades a priori, y los costes para tomar una decisión. En concreto, continuando con el ejemplo de los vidrios. El papel del forense se limitará a calcular la razón de verosimilitudes entre el vidrio de la ventana y el vidrio de la chaqueta del sospechoso. El juez, será el encargado de tomar una decisión a partir de probabilidades a priori, y teniendo en cuenta los costes de decisión. Es decir, por el sistema penal de presunción de inocencia en España, tiene un mayor coste asociado el incriminar a un inocente que el salvar a un culpable.

Suponiendo que se tiene una evidencia E y una hipótesis θ binaria. Se tendrá θ_p cuando las fuentes dubitadas e indubitadas coincidan, en nuestro caso forense se dirá que se tienen dos marcas de la **misma fuente**. Y, se tendrá θ_d cuando no coincidan, es decir, cuando las dos marcas sean de **fuentes distintas**. [1] Siguiendo con el ejemplo planteado de los vidrios, se dirá θ_p cuando los vidrios de la chaqueta del sospechoso y de la ventana coincidan y θ_d cuando no coincidan.

Utilizando el Teorema de Bayes [10] se puede calcular la probabilidad de nuestra hipótesis teniendo en cuenta la evidencia, en concreto, se puede calcular por ejemplo la probabilidad de que nuestros dos vidrios provengan de la misma fuente dadas unas evidencias o indicios:

$$P(\theta_p|E) = \frac{P(E|\theta_p) \times P(\theta_p)}{P(E)} \tag{2.1}$$

siendo $P(E|\theta_p)$ la verosimilitud y $P(\theta_p)$ la probabilidad a priori de que ambos vidrios provengan de la misma fuente sin haber examinado las evidencias.

Así, relacionando la probabilidad de que las muestras vengan de la misma fuente con la probabilidad de que vengan de distintas fuentes, la ecuación resultante es la siguiente:

$$\frac{P(\theta_p|E)}{P(\theta_d|E)} = \frac{P(\theta_p)}{P(\theta_d)} \times \frac{P(E|\theta_p)}{P(E|\theta_d)} \tag{2.2}$$

En esta ecuación, en el miembro de la izquierda del igual se relacionan las **probabilidades a posteriori**, es decir, la probabilidad predicha de que la clase sea una u otra. Siguiendo con nuestro ejemplo, se relacionan las probabilidades de que los vidrios provengan de la misma fuente y de distintas fuentes.

En el primer miembro de la derecha del igual, tendremos la relación entre las probabilidades **"a priori"**, las cuáles no son responsabilidad del forense. Como hemos visto en el ejemplo anterior, el forense tiene que limitarse a dar una relación de verosimilitudes sin tener en cuenta ninguna probabilidad a priori. Es el juez, el que utilizará esa información a priori para tomar una decisión.

Además, ya hemos visto que otra cosa que tenía en cuenta el juez para tomar una decisión son los costes asociados a cada decisión incorrecta. Llamaremos C_{fr} al coste de predecir θ_d cuando θ_p, es decir, será el coste de predecir que los vidrios pertenecen a fuentes distintas cuando no lo son (salvaremos al culpable). Por el contrario, tendremos C_{fa} en el caso contrario, cuando predecimos que los vidrios son la misma fuente y en realidad no lo son (incriminaremos al inocente).

Volviendo a la ecuación 2.2, en el último miembro, tenemos la **razón de verosimilitud** entre las dos hipótesis. A este ratio de similitues también se le llama **LR**, el cuál será el asignado por el sistema forense, y por tanto, será el que estudiaremos un poco más a fondo en este TFG. [11]

Además, debido a que se tienen dos clases complementarias sabemos:

$$P(\theta_p) = 1 - P(\theta_d) \tag{2.3}$$

Ahora, podremos reflejar la relación de probabilidades a priori y a posteriori de la siguiente manera solo en función de una de las hipótesis:

$$O(\theta_p) = \frac{P(\theta_p)}{P(\theta_d)} = \frac{P(\theta_p)}{1 - P(\theta_p)} \tag{2.4}$$

$$O(\theta_p|E) = \frac{P(\theta_p|E)}{P(\theta_d|E)} = \frac{P(\theta_p|E)}{1 - P(\theta_p|E)} \tag{2.5}$$

Por tanto, introduciendo 2.4 y 2.5 en 2.2, tendremos la ecuación resultante:

$$O(\theta_p|E) = LR \times O(\theta_p) \tag{2.6}$$

Finalmente, juntando 2.3 y 2.6 podemos obtener la expresión para la probabilidad a posteriori de una clase:

$$P(\theta_p|E) = \frac{LR \times O(\theta_p)}{1 + LR \times O(\theta_p)} \tag{2.7}$$

Ya hemos visto que el LR es una relación de verosimilitudes del siguiente estilo:

$$LR = \frac{P(E|\theta_p)}{P(E|\theta_d)} \tag{2.8}$$

Suponiendo que ambos costes, C_{fr} y C_{fa}, son iguales, la regla de decisión de Bayes dicta que se decidirá θ_p cuando $P(\theta_p|E) > P(\theta_d|E)$, y se decidirá θ_d cuando $P(\theta_d|E, I) > P(\theta_p|E, I)$. Se puede demostrar que, en este caso, la regla de Bayes minimiza la probabilidad de error [12]. Así, podemos describir los dos tipos de error:

- **Falso negativo (FR)**: Es el error al predecir $\theta = \theta_d$ cuando θ_p es cierta, es decir, se predicen clases diferentes cuando son la misma. C_{fr} será el coste asociado a dicha equivocación.

- **Falso positivo (FA)**: Es el error al predecir $\theta = \theta_p$ cuando θ_d es cierta, es decir, se predicen clases iguales cuando son diferentes. C_{fa} será el coste asociado a éste error.

Mediante la información a priori y los costes de decisión explicados se establece una frontera de decisión, que en realidad es un umbral escalar, llamado umbral de bayes. De manera que, si el LR está por encima de dicha frontera se decide una clase y si está por debajo se decide otra. Se define como sigue:

$$\tau = \frac{P(\theta_d|I) \times Cfa}{P(\theta_p|I) \times Cfr} \tag{2.9}$$

Estableciendo el **umbral de Bayes (B)**, las dos curvas de verosimilitudes θ_p y θ_d y las dos regiones R_d y R_p donde prediceremos $\theta = \theta_d$ y $\theta = \theta_p$ respectivamente, podemos definir los valores de falso negativo y falso positivo. Un ejemplo lo podemos ver en la imagen 2.3.

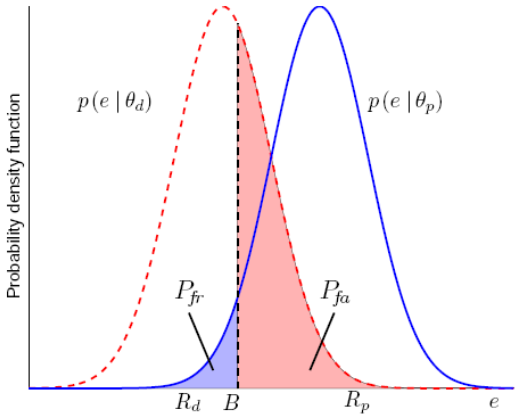

Figura 2.3: *Ejemplo del umbral de Bayes(B) con la probabilidad de falso negativo (Pfr) y la probabilidad de falso positivo (Pfa) [1]*

Una vez explicado el modelo bayesiano es importante que éste se comporte de la forma en la que debería. Una posible medida del rendimiento de una probabilidad a posteriori es la desviación de dicho valor probabilístico del caso en el que θ_p sea cierta ($P(\theta_p|E) = 1$) o de que sea falsa ($P(\theta_p|E) = 0$).

Una medida de desviación que utilizaremos serán las llamadas **SPSR (Strictly Proper Scoring Rules)** [5]. En definitiva, lo que intenta esta medida es asignar una penalización cuánto más lejos de la verdad nos encontremos. Si lo pensamos este método tiene sentido. *No es lo mismo que te digan que habrá una probabilidad del 60 % de lluvias mañana y que no llueva, a que te afirmen al 100 % que mañana llueve y se equivoquen. En este último caso, se le asignará una penalización mucho mayor al hombre del tiempo, pues a la próxima vez probablemente no confiaríamos en lo que nos dice.*

Esto lo podemos ver representado en la gráfica de la imagen 2.4. La curva roja representa cuando $\theta = \theta_p$, y la curva azul cuando $\theta = \theta_d$. Así, diremos que cuanto mayor es la probabilidad a posteriori de que las dos muestras sean iguales ($P(\theta_p|E)$), la curva roja tenderá a 0, es decir, cuanto más cerca de la verdad estemos menor será la penalización. En el caso que la verdad sea que son de fuentes distintas(curva azul), a medida que aumentemos el valor de $P(\theta_p|E)$ más nos estaremos alejando de la verdad, por lo tanto, iremos aumentando la penalización considerablemente.

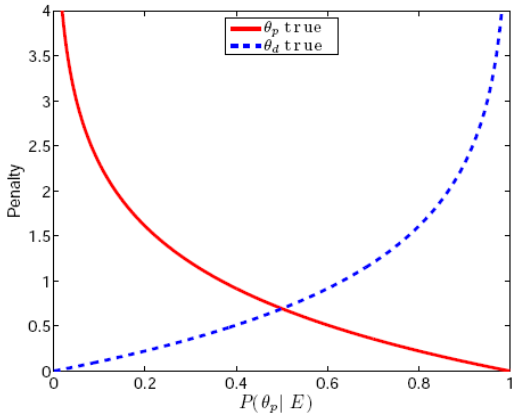

Figura 2.4: *Regla de puntuación logarítmica* [1]

Como hemos visto en la imagen anterior la curva roja sigue una distribución logarítmica, es decir, cuando $\theta = \theta_p$, la penalización se puede definir como $-\log_2 P(\theta_p|E)$.

Imaginemos que queremos evaluar la exactitud del hombre del tiempo. En ese caso, podremos analizar varias de las predicciones hechas no solo una. [13]. Así, obtenemos conjuntos de probabilidades y podemos definir SPSR como la siguiente ecuación:

$$LS = -\frac{1}{Np} \cdot \sum_{\theta^i=\theta_p} \log_2 P(\theta^i|E) - \frac{1}{Nd} \cdot \sum_{\theta^j=\theta_d} \log_2 P(\theta^j|E) \tag{2.10}$$

siendo N_p y N_d el número de comparaciones para las que θ_p y θ_d son ciertas respectivamente en el conjunto experimental, y que representan por tanto, comparaciones entre muestras de la misma fuente y de fuentes diferentes.

2.3. Proceso de validación forense

Un proceso forense está compuesto generalmente por 4 tipos de actividades: la escena del crimen, el análisis, la interpretación y el reportaje. En el contexto de interpretación de las evidencias será donde tendremos la implementación del método y la validación de este. Así, podemos definir **validación** como *el proceso necesario para determinar el grado de validez de un método dado para calcular los LRs.* [14]

El desarrollo de los métodos para calcular los LRs tiene una combinación de varias labores:

- La primera labor se centra en la metodología forense en la que los científicos forenses intentan **desarrollar nuevos métodos y soluciones** a partir de preguntas que necesitan ser respondidas.
- La segunda, se centra en la **implementación y validación de esos nuevos métodos y soluciones**. Con esto, intentan encontrar el alcance de validez cuando esas metodologías se usan en un entorno forense. Lo hacen simulando la funcionalidad del método en un amplio rango de condiciones y resultados que permita establecer límites hasta los cuáles haya una confianza. A continuación hablaremos un poco más de esta labor.
- La tercera y última labor se centra en la **evaluación forense en la práctica**. Es decir, se usan nuevos métodos y soluciones en un caso real para evaluar la solidez de una evidencia con respecto a proposiciones alternativas.

La segunda labor es importante. El proceso de validación es aquel que nos indica el grado de confianza que podemos depositar en ese método en un caso real e independientemente del dataset ante el que nos encontremos. Es por eso que **el conjunto de datos** utilizado para la validación debe presentar condiciones forenses reales. Además, debería ser independiente y representativo con respecto al conjunto de datos de la época de desarrollo.

A continuación, explicaremos un poco más a fondo la segunda labor la cuál se divide en dos etapas:

- **La etapa de desarrollo del método**: aquí proponemos tratar con procesos relacionados con la selección del método, la fase de entrenamiento o la fase de test, y medir las características de rendimiento primario tales como precisión, calibración o poder discriminativo. (Ver en el apartado 2.4).

- **La etapa de validación del método**: en esta etapa evaluaremos el desempeño del método de LRs utilizando un conjunto de datos de validación y mediremos el rendimiento tanto de las características primarias como de las características secundarias, tales como, coherencia, generalización y robustez [14]. En este TFG en concreto, la medida de la que hablaremos más a fondo será la **robustez** [definida más adelante en el apartado 3.3].

En este apartado es importante distinguir entre los procesos de desarrollo y validación en el campo forense y en el campo de inteligencia artificial (Machine Learning). Esta comparación se puede ver claramente en la figura 2.5.

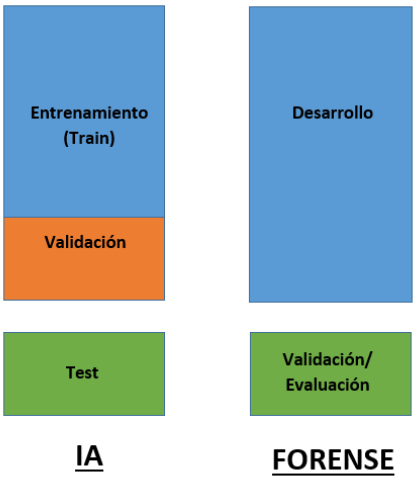

Figura 2.5: *Esquema de desarrollo y validación de métodos en distintos campos: el forense y en inteligencia artificial.*

Como podemos ver en la figura anterior, hay determinados conceptos que no significan lo mismo para uno y otro campo. En inteligencia artificial separamos el proceso en la fase de entrenamiento del modelo y la de validación o ajustación de modelo. En forense, sin embargo, estas dos fases estarán juntas en el denominado proceso de desarrollo. Además, en inteligencia artificial se llama fase de test a la fase en la que se prueba el modelo en un entorno real. Sin embargo, en forense, esta fase en la que se prueba un modelo realista es la llamada Validación. Es importante tener en cuenta esto, porque como podemos ver se llama Validación a fases diferentes en campos distintos.

2.4. Precisión: Poder Discriminativo y Calibración

Es importante que un modelo sea exacto, es decir, que se ajuste todo lo posible al valor real. Esto es, siguiendo con los términos de probabilidades descritas en el apartado anterior, que asigne $P(\theta_p|E) = 1$ si $\theta = \theta_p$ y, $P(\theta_p|E) = 0$ si $\theta = \theta_d$.

Dos de las características más importantes que necesita un sistema para que sea exacto serán: el poder discriminativo y la calibración [15] [1]

Poder discriminativo

Se trata de una medida que evalúa el **grado en el que el modelo distingue una hipótesis de otra**, es decir, diremos que un modelo tiene una buena discriminación si se tiene la capacidad

de separar correctamente las evidencias de la misma fuente y de fuentes distintas. En términos generales podemos decir que se trata de la habilidad para dar información sobre el valor de la hipótesis.

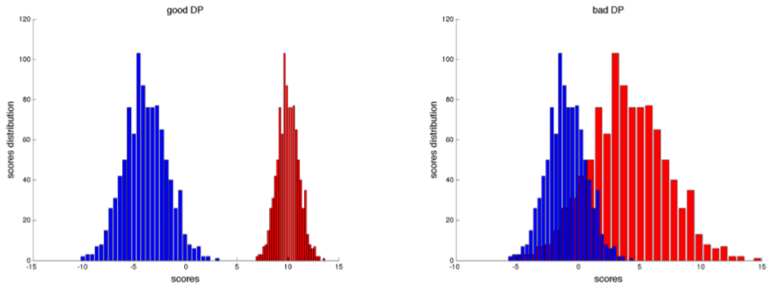

Figura 2.6: *Ejemplos de gráficos con buena y mala discriminación. [3]*

En la imagen 2.6 podemos ver dos gráficas en las que el poder discriminativo es muy diferente. Mientras en la primera tendremos un alto poder discriminativo pues podemos diferenciar bien las dos clases, en la segunda hay un momento que es difícil distinguir de qué clase es un valor. En la primera, podríamos decir que todos los scores menores a 5 pertenecerían la clase azul y los mayores, a la clase roja. En la segunda, sin embargo, hacer esta distinción es más complicado, surgiendo así los **falsos positivos** y **falsos negativos**. Esto es, cuando predecimos un valor como clase azul cuando es en realidad roja, o al contrario.

Calibración

Es la medida que compara la probabilidad predictiva de la variable que se quiere predecir o **probabilidad a posteriori"** con la proporción real de casos donde se observa la variable a predecir. Cuánto más semejantes sean, el modelo más calibrado estará. Un ejemplo de esto sería: *dadas dos clases, población enferma y población sana, si el modelo predice que la población enferma es del 20 % estará perfectamente calibrado si, en efecto, la población enferma es del 20 %.* [16]

Una manera de medir la exactitud es mediante el **SPSR**. El SPSR mide la desviación con respecto a la verdad. Así, un sistema perfecto tendría un SPSR = 0. Puesto que la exactitud se puede medir en función de SPSR y hemos dicho que las medidas más importantes para que un sistema sea exacto son el poder discriminante y la calibración, podemos definir el SPSR y así, la exactitud, en términos de estas dos medidas de rendimiento:

$$C = C^{cal} + C^{disc} \tag{2.11}$$

Siendo **C**, la media del coste debido a la falta de exactitud, entonces ésta dependerá del coste de falta de calibración o C^{cal} y del coste de la falta del poder discriminativo o C^{disc}. Así, podemos decir que necesitamos un sistema equilibrado en el que tanto calibración como poder discriminativo vayan de la mano. [17]

Para probar lo exacto que es un modelo utilizaremos las llamadas **medidas de rendimiento**. Una medida de rendimiento en la que profundizaremos en este TFG serán **las curvas ECE**.

3

Diseño

3.1. Teoría de la información: Entropia

La teoría de la información afirma que la información obtenida en un proceso inferencial es determinado por la reducción de la entropía, *la cual mide la incertidumbre sobre una determinada variable respecto a una información conocida.* Esto significa que cuanto mayor sea la entropía, mayor será la incertidumbre. En el marco forense, se utiliza la entropía para representar la incertidumbre que existe en cada caso acerca del valor verdadero de la hipótesis. [18]

En un problema de interpretación forense, la incertidumbre sobre la hipótesis se da a partir de las probabilidades a priori. Es por eso que la entropía también puede llamarse **"Entropía a priori"** cuando no se ha observado la evidencia y se da con la siguiente fórmula:

$$H_p(\theta) = -\sum_{i=p,d} P(\theta_i) \times log_2 P(\theta_i) \tag{3.1}$$

En la imagen 3.1 podemos ver la curva de la entropía para $P(\theta_p)$. Podemos ver que la entropía es máxima cuando $P(\theta_p) = 0{,}5$, y por lo tanto, cuando $P(\theta_d) = 0{,}5$ [4]. Esto tiene sentido, puesto que si lo pensamos tener una probabilidad del 50 % es la máxima incertidumbre que podemos tener, es imposible predecir si se trata de una hipótesis u otra, es decir, si $\theta = \theta_p$ o $\theta = \theta_d$.

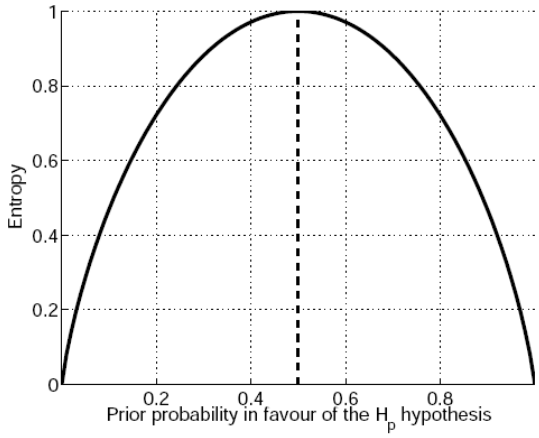

Figura 3.1: *Entropía a priori con respecto a* $P(\theta_p)$. *[1]*

Además, se puede demostrar que el valor esperado de la entropía de las probilidades a posteriori sobre todos los valores de la evidencia E, es siempre menor que el valor de la entropía a priori, [4] y, se puede expresar con la ecuación 3.2.

$$H_p(\theta|E) = -\sum_{i=p,d} P(\theta_i) \int_e P(e|\theta_i) log_2 P(\theta_i|e)\, de \tag{3.2}$$

Además en términos de la entropía a posteriori hay que tener en cuenta una divergencia entre el mejor valor con el mejor modelo y los valores calculados con el resto de modelos. Ya hemos visto que dependiendo del modelo que utilicemos nos aproximaremos más al valor real de la hipótesis o no. Es por eso, que definiremos la entropía cruzada como una descomposición de la entropía a posteriori, obtenida a partir de las entropías a priori y de la evidencia, más la divergencia entre el mejor valor y el resto de valores posibles:

$$H_{p||\tilde{p}}(\theta|E) = H_p(\theta|E) + D_{P||\hat{P}}(\theta|E) \tag{3.3}$$

En la figura 3.2 mostraremos la pérdida de información medida por la entropía cruzada en términos de su descomposición.

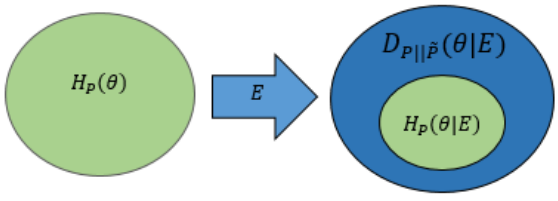

Figura 3.2: *Esquema de la descomposición de la entropía cruzada. Adaptado de [4]*

La figura 3.2 muestra un esquema de cómo con la entropía de referencia y con la evidencia, podemos obtener información acerca de la hipótesis (entropía a posteriori). La divergencia que es el óvalo azul, muestra la pérdida de información adicional al utilizar un mal modelo, siendo la menor pérdida representada en el óvalo verde que está en su interior.

3.2. Medidas de rendimiento: Curvas ECE y C_{llr}

Las curvas ECE (**Empirical Cross Entropy**) son *una manera de representar el rendimiento de un conjunto de valores LR midiendo la precisión en términos de calibración y poder discriminativo.* [5]

Hay varias maneras de definir las curvas ECE, se pueden ver como una aproximación de la entropía cruzada $H_{p||\hat{p}}(\theta|E)$ [4] o como la media del valor de la regla de puntuación o **SPSR**:

$$ECE = -\frac{P(\theta_p|I)}{N_p}\sum_{\theta=\theta_p} log_2 P(\theta_p|E_i, I) - \frac{P(\theta_d|I)}{N_d}\sum_{\theta=\theta_d} log_2 P(\theta_d|E_i, I) \tag{3.4}$$

A partir de la ecuación 3.4 y con lo visto en el apartado 2.2 de inferencia bayesiana, podemos mostrar la expresión de la curva ECE en términos de la relación de probabilidades a priori y de los valores LR:

$$ECE = \frac{P(\theta_p|I)}{N_p}\sum_{\theta=\theta_p} log_2(1 + \frac{1}{LR_i \times O(\theta_p)}) + \frac{P(\theta_d|I)}{N_d}\sum_{\theta=\theta_d} log_2(1 + LR_i \times O(\theta_p)) \tag{3.5}$$

Las curvas ECE se suelen representar como funciones logarítmicas de las probabilidades a priori, es decir, variaremos el valor de las probabilidades a priori, puesto que en muchos casos son variables desconocidas para el forense, y así obtendremos su curva ECE. Un ejemplo lo vemos en la figura 3.3.

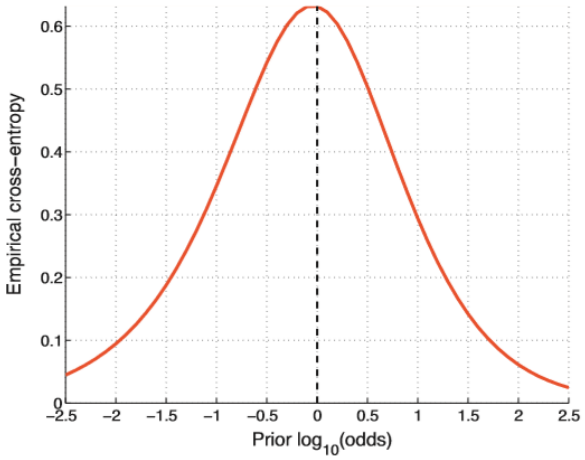

Figura 3.3: *Curva ECE como función logarítmica de las probabilidades a priori. [5]*

Sin embargo, en este tipo de gráficas se suelen representar más curvas que nos den más información sobre la precisión. Los tres tipos de curvas a representar son los siguientes:

- **Curva sólida de color rojo**: Esta es la curva mostrada en la figura 3.3. Muestra la entropía cruzada, es decir, la pérdida media de información de los valores LR calculados. Cuanto menor es esta curva, mejor es la exactitud, puesto que hay menos pérdida de información. Teniendo en cuenta la figura 3.2, esta curva será la correspondiente con el óvalo azul, es decir, es la entropía del modelo utilizado (sin tener por qué ser la mejor entropía).

- **Curva azul discontinua**: Esta es la curva que se obtiene al aplicar el algoritmo PAV a los valores LR calculados. Se puede demostrar que el **agoritmo PAV** (Pool Adjacent Violators) implementa una técnica para optimizar *Cllr* preservando el poder discriminativo del conjunto de scores original [19] [20]. Es decir, se calcula el límite inferior al que el sistema puede llegar para mejorar su calibración sin modificar el poder de discriminación. Esta curva mide el poder de discriminación, de manera que, cuanto menor sea la curva, mayor poder de discriminación. Puesto que es la mejor entropía que se puede obtener, diremos que coincide con el óvalo verde pequeño de la derecha en la figura 3.2.

- **Curva negra punteada**: Esta curva es una referencia neutral, representando el rendimiento de un sistema con un $LR = 1$ siempre. Es importante que, esta curva esté siempre por encima de la curva sólida, de lo contrario el método de cálculo de LR será peor que asignar LR=1 en todos los casos, es decir, peor que no hacer nada.

Además, también se podrá medir la calibración a partir de estas curvas, de manera que, cuánto más cercanas estén las curvas roja y azul, mejor será la calibración.

A continuación, veremos dos ejemplos donde se muestran las medidas de rendimiento explicadas. En la imagen 3.4, podemos ver que la curva azul es bastante pequeña , es decir, se aleja bastante de la curva neutral o curva negra punteada. Además, está bastante cerca de la roja, teniendo así buena calibración y buen poder discriminativo. En la imagen 3.5, por el contrario, aunque aparecen las dos curvas roja y azul muy cerca teniendo así buena calibración, éstas son bastante grandes, es decir, se acercan bastante a la curva negra neutral, perdiendo así poder discriminativo.

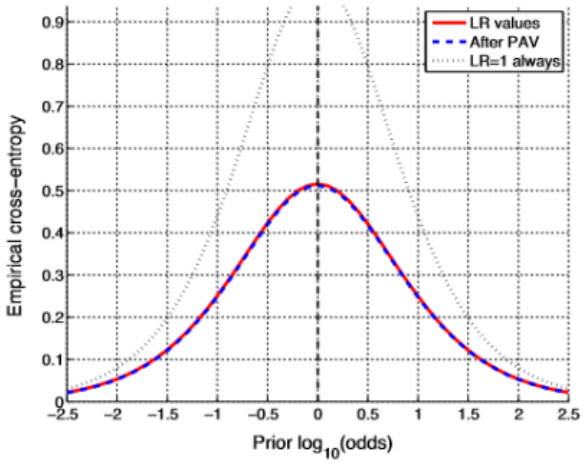

Figura 3.4: *Curvas ECE de un ejemplo bien calibrado y con buen poder discriminativo. [1]*

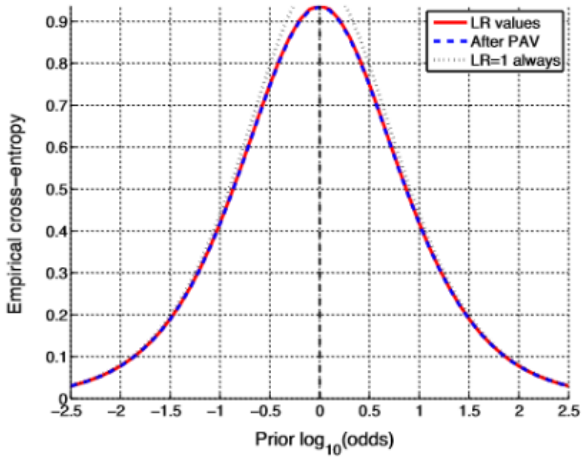

Figura 3.5: *Curvas ECE de un ejemplo bien calibrado y con mal poder discriminativo. [1]*

Ya hemos visto que ECE es una aproximación de la entropía cruzada ($H_{p||\hat{p}}(\theta|E)$). Por lo tanto, se podrá descomponer como tal de la siguiente manera:

$$ECE = ECE^{min} + ECE^{cal} \tag{3.6}$$

siendo ECE^{min} una aproximación de $H_p(\theta|E)$ que representa la información perdida por la falta de discriminación y ECE^{cal} una aproximación de $D_{P||\hat{P}}(\theta|E)$ que representa la información perdida por la falta de calibración.

Una vez definidas y entendidas las curvas ECE podemos definir el C_{llr} como el valor de ECE donde las probabilidades a priori se igualan, es decir, *es el coste de decisión esperado para cualquier valor de* C_{fa} *y* C_{fr} *y para* $P(\theta_p) = P(\theta_d) = 0{,}5$. Así, a partir de la regla de puntuación, podemos proponer la siguiente ecuación para medir la precisión en ese punto:

$$C_{llr} = \frac{1}{2 \cdot N_p} \sum_{i=p} log_2(1 + \frac{1}{LR_i}) + \frac{1}{2 \cdot N_d} \sum_{j=d} log_2(1 + LR_j) \tag{3.7}$$

siendo N_p y N_d el número de comparaciones (valores LR) para las que θ_p y θ_d son ciertas respectivamente, y que, representan por tanto, las comparaciones de fuentes iguales y de fuentes diferentes.

Podemos descomponer el C_{llr} igual que las curvas ECE en la fórmula 3.6: en C_{llr} asociado al poder discriminativo (C_{llr}^{min}) y C_{llr} asociado a la calibración (C_{llr}^{cal}):

$$C_{llr} = C_{llr}^{min} + C_{llr}^{cal} \tag{3.8}$$

Así, las curvas ECE se pueden definir también como una generalización del C_{llr} permitiendo un análisis más detallado del método de los LR. Sin embargo, el C_{llr} presenta la ventaja de ser un valor escalar, y por lo tanto ser útil como medida única resumida de rendimiento, y como resumen de una curva ECE. Ambas medidas de rendimiento se utilizarán en este TFG.

3.3. Robustez

Además de las medidas de rendimiento primario tales como calibración y poder discriminativo, habíamos hablado de las medidas de rendimiento secundario como coherencia, generalización y robustez. En este apartado hablaremos un poco más a fondo de una de esas medidas de rendimiento secundario, **la robustez**, puesto que será utilizada en este TFG. [14]

La robustez en concreto, es una medida de la capacidad del modelo para permanecer no afectado por variaciones en parámetros del método, proporcionando así, una indicación de su fiabilidad durante su uso normal. [21]

Un ejemplo de variación en los parámetros podría ser el cambio en la calidad de los datos o en su cantidad. Es entendible que con más cantidad de datos, el resultado fácilmente sea más próximo al real, y con datos peores que nos aporten poca información, el resultado se alejará del real. Por lo tanto, diremos que un sistema es más robusto que otro si, al variar la calidad de los datos y disminuir su cantidad, la medida de rendimiento se degrada relativamente poco.

Podemos definir que **la robustez en el contexto LR** usualmente se refiere a la estabilidad de los métodos de LRs al variar las condiciones (*por ejemplo, calidad / cantidad de los datos*). En el caso de que dicha degradación venga definida por la incertidumbre del modelo debida a la escasez de los datos, dicha estabilidad se puede medir con **intervalos de confianza** de la hipótesis estudiada. Definimos intervalo de confianza [22] como un par de números entre los cuales se estima que estará el rendimiento (o exactitud) de nuestro método de cálculo de LR con una determinada probabilidad de acierto. Formalmente, estos números determinan un intervalo, que se calcula a partir de datos de una muestra.

4

Desarrollo

Tal y como hemos explicado anteriormente, este TFG se centra en el estudio de cómo probar la robustez de nuestro modelo, esto es, generando unos intervalos de confianza en los que se encuentre nuestro valor con un determinado porcentaje de acierto, y, es por eso, que en este apartado explicaremos algunas maneras de crearlos. [21]

4.1. Métodos paramétricos de cálculo de intervalos de confianza

La técnica más básica para poder medir la robustez mediante intervalos de confianza es la llamada técnica paramétrica, que consiste en una técnica de distribución de muestreo. Así, surgen dos definiciones importantes: muestra y muestreo. [23]

Una **muestra** es una porción representativa de elementos de una población, elegida para un experimento. Como nos podemos imaginar resulta costoso el análisis de todos los datos, así que se hace solo sobre una porción representativa de la población.

Muestreo será por lo tanto, el proceso de selección de muestras que se utiliza cuando no nos es posible contar o medir todos los elementos de la población. [24]

Uno de los parámetros sobre los que se puede y se suele hacer muestreo, es la media o esperanza de la población. Así, el experimento teórico podría consistir en obtener múltiples muestras de tamaño N de una muestra inicial para calcular la media de cada muestra, y poder así, representar la distribución de las medias obtenidas. Además, hay que destacar que la desviación estándar será el **error estándar** que nos permitirá calcular el intervalo de confianza que deseemos.

Por el Teorema del Límite Central [6] que nos permite asumir gaussianidad y suponiendo que conocemos la varianza y que nuestros datos son independientes, podemos asegurar que si la muestra es grande la distribución de la media muestral sigue una distribución normal.

Puesto que en el entorno forense, no sabemos la varianza de la muestra, tendremos que estimarla, siguiendo así otra distribución llamada **T-student.** Hay que tener en cuenta que cuanto mayor sea N, mejor estimada estará la varianza por lo que más nos aproximaremos a una normal.

Ya hemos dicho que lo que estamos estudiando en este trabajo son los márgenes de error, por lo que buscaremos calcularlos a partir de esta distribución. En la figura 4.1 mostramos la distribución t-student y cómo coger los intervalos de confianza. En términos generales si queremos un intervalo de int $= (1-\alpha)100\,\%$, entonces se cogerán los valores de los percentiles $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$. Por ejemplo, para tener un intervalo del 90 % de acierto, el procedimiento consiste en quitar a esta distribución el primer 5 % y el último (haremos los percentiles 5 % y 95 %). Tener un intervalo del 90 % significará que el valor que buscamos estará entre esos valores con una probabilidad del 90 %.

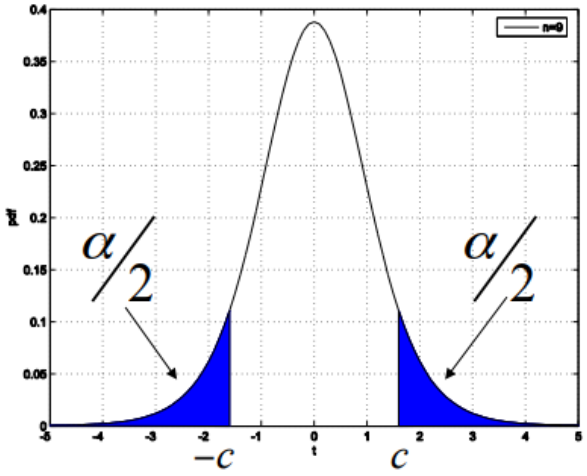

Figura 4.1: *Distribución t-student. [6]*

4.2. Técnica Bootstrap de cálculo de intervalos de confianza

El procedimiento de **Bootstrapping** permite inferir la distribución de una variable estadística a partir de un conjunto de muestras de una población cuya distribución no se conoce. Consiste en un método de muestreo con reposición, es decir, dado un conjunto de N elementos, una muestra bootstrap consiste en extraer N elementos con reemplazamiento. Una característica importante es que bootstrap siempre supone que los **elementos son independientes**.

En la figura 4.2 podemos ver como sería coger una muestra con reposición. Sin embargo, hay que tener en cuenta que en una muestra bootstrap este proceso de sacar muestras con reposición se repetirá un gran número de veces, obteniendo así, un alto número de estimaciones.

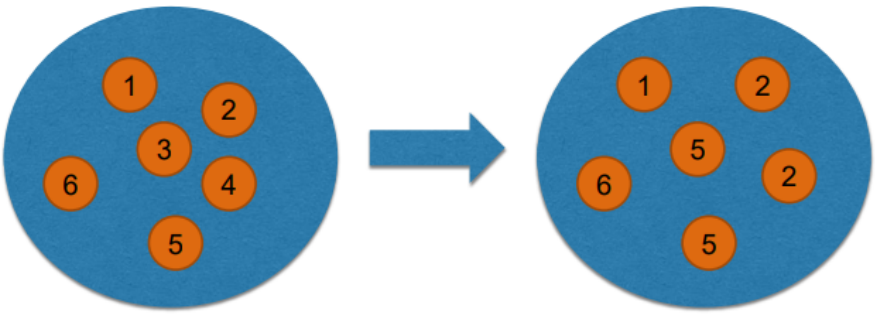

Figura 4.2: *Ejemplo de una muestra bootstrap.[7]*

4.2.1. Bootstrap aplicado a la ciencia forense

La técnica bootstrap se puede utilizar para clasificar hipótesis en un entorno concreto, por ejemplo el forense. El objetivo es el mismo, lo que nos interesa es crear intervalos de confianza tal y como hemos visto en el apartado anterior para estimar parejas de características biométricas. En particular, lo que nos interesa es estimar si esas parejas pertenecen a la misma fuente o a fuentes distintas.

Tendremos un conjunto de pares de la misma fuente $M = a_1, a_2, ..., a_m$ y otro conjunto de pares de distintas fuentes $N = b_1, b_2, ..., b_n$. Además, para emparejar a esos pares en una misma fuente o en fuentes distintas, será necesario conocer el valor de sus LRs. Siendo X los LRs de los pares de la misma fuente e Y los LRs de pares de fuentes distintas tendremos $X = x_1, x_2, ..., x_m$ e $Y = y_1, y_2, ..., y_n$. Donde M < N puesto que siempre encontraremos más pares de fuentes distintas que de la misma fuente. [21]

Aunque las técnicas bootstrap se han aplicado en el pasado para la estimación de intervalos de confianza en tasas de error [21], en el entorno forense, no tenemos una decisión a partir de la cuál hallar estos errores, lo que intentamos es aportar el valor de la prueba en forma de LR, por lo que el método a seguir para calcular los intervalos de confianza, serán las anteriormente explicadas curvas ECE.

A continuación mostraremos el proceso de una muestra bootstrap:

1. **Calculamos la curva ECE**: se usará la ecuación 3.5 para calcular la curva ECE a partir de los LRs de la misma fuente (X) y de fuentes distintas (Y).

2. **Reemplazamiento**: se crea la muestra bootstrap con reemplazamiento a partir de la muestra inicial X e Y.

3. **Estimación**: se calculan los nuevos valores de una curva ECE con la muestra obtenida en el paso anterior.

4. **Repetición**: se repiten los pasos 2 y 3 un número elevado de veces (B).

Para determinar los intervalos de confianza, una vez teniendo todas las curvas ECE de todas las repeticiones, las ordenamos en orden ascendente, pudiendo representarlas en un histograma, y determinaremos el intervalo de confianza.

Así, diremos que nuestro intervalo del $(1-\alpha)100\,\%$ será $[ECE_{k1}, ECE_{k2}]$, siendo $k_1 = \frac{\alpha}{2} \cdot B$ y $k_2 = (1 - \frac{\alpha}{2}) \cdot B$. Por ejemplo, *para conseguir un intervalo del 90 %, cogeremos el intervalo entre los valores de los percentiles 5 % y 95 %.* Ver figura 4.3.

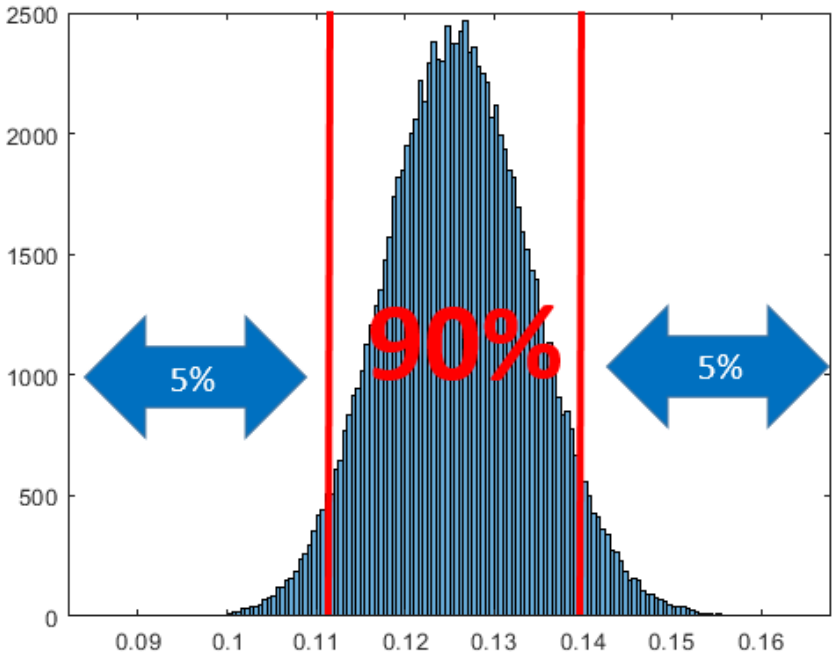

Figura 4.3: *Ejemplo de los intervalos de confianza de una muestra bootstrap. [8]*

Sin embargo, la técnica de Bootstrapping considera que todos los scores son independientes. En el caso de que esto no sea así, no nos dará unos intervalos de confianza fiables, es decir, nos dará unos intervalos de confianza menores que los que nos debería dar. Esto es debido a que subestima estos intervalos al considerar todos los datos independientes cuando no lo son.

4.3. Técnica Subset Bootstrap de cálculo de intervalos de confianza

Como hemos dicho antes, a veces por la forma en la que los datos están recogidos, es bastante probable la dependencia entre los valores. Por ejemplo, en nuestra base de datos de vidrios, el LR obtenido de la comparación del vidrio 1 con el vidrio 2 y el LR del vidrio 1 con el vidrio 3, podemos decir que son dependientes, pues ambos LRs dependen del vidrio 1.

Es por eso que necesitamos otro método para hallar los intervalos de confianza cuando hay dependencia entre los datos. El que vamos a estudiar en este TFG es el **Subset Bootstrap**.

La idea de este método es dividir el conjunto de muestras en subconjuntos, de manera que, dentro de cada subconjunto las muestras tengan dependencia entre ellas, pero, independencia con el resto de muestras de otros subconjuntos.

4.3.1. Subset Bootstrap aplicado a la ciencia forense

Teniendo en cuenta la nomenclatura del apartado 4.2.1, es decir, siendo $X = x_1, x_2, ..., x_m$ los LRs de los pares de muestras de la misma fuente e $Y = y_1, y_2, ..., y_n$ los LRs de pares de muestras de fuentes diferentes, y suponiendo que los LRs son dependientes unos de otros, haremos subconjuntos.

Vamos a suponer que el conjunto de LRs de las muestras son dependientes unos de otros. En ese caso nuestra labor será crear subconjuntos de LRs que sean dependientes entre sí, pero que a la vez, los LRs de cada subconjunto sean independientes con los LRs del resto de subconjuntos. Así, suponiendo que tenemos D subconjuntos, podremos volver a nombrar a los LRs de pares de la misma fuente como $X = x_{11}, x_{12}, ..., x_{21}, x_{22}, ..., x_{1D}, x_{2D}, ...$ y los de distinta fuente como $Y = y_{11}, y_{12}, ..., y_{1D}, y_{2D}, ...$ donde el primer subíndice indicará el número del LR de la misma fuente o de fuentes distintas dentro del subconjunto y el segundo subíndice indicará el subconjunto al que pertenece. Podemos decir entonces, que todos los que tienen un segundo subíndice igual serán dependientes entre ellos, e independientes con el resto. [21]

A continuación mostraremos el proceso de una muestra subset bootstrap:

1. **Calculamos la curva ECE**: se usará la ecuación 3.5 para calcular la curva ECE a partir de los LRs de la misma fuente (X) y de fuentes distintas (Y).
2. **Reemplazamiento**: se crea la muestra bootstrap con reemplazamiento de cada subconjunto, es decir, a partir de los subconjuntos creados anteriormente, se coge una muestra considerando cada subconjunto como un solo valor que identifique al subconjunto, de manera que o se coge el subconjunto entero o no se coge.
3. **Estimación**: se calculan los nuevos valores de una curva ECE con la muestra obtenida en el paso anterior.
4. **Repetición**: se repiten los pasos 2 y 3 un número elevado de veces (B).

Los intervalos de confianza se determinarán de la misma manera que con la muestra bootstrap: una vez teniendo todas las curvas ECE, las ordenamos en orden ascendente, y diremos que nuestro intervalo del $(1-\alpha)100\,\%$ será $[ECE_{k1}, ECE_{k2}]$, siendo $k_1 = \frac{\alpha}{2} \cdot B$ y $k_2 = (1 - \frac{\alpha}{2}) \cdot B$.

Para comparar las dos técnicas descritas, consideremos un ejemplo de dos variables aleatorias X_1 y X_2 con media cero. [21] Además, sabemos que la varianza o también llamado **momento centrado de orden 2** es una medida de dispersión y se puede definir con la siguiente expresión:

$$\sigma^2 = E[(X - \mu)^2] \tag{4.1}$$

Si suponemos que ambas variables aleatorias son independientes, entonces sabemos que la varianza de la suma es igual a la suma de sus varianzas. Y, puesto que las dos variables están igualmente distribuidas, podemos decir que su varianza será:

$$E[(X_1 + X_2)^2] = 2E[(X1)^2] \tag{4.2}$$

Si por el contrario, ambas variables fueran dependientes, y tomamos la máxima dependencia, es decir, suponemos que ambas son iguales, la varianza de su suma sería la siguiente:

$$E[(X_1 + X_2)^2] = 4E[(X1)^2] \tag{4.3}$$

Como vemos, la varianza de variables dependientes es siempre mucho mayor a la varianza de variables independientes. Por lo tanto, en el caso de que los datos sean dependientes unos de otros, con la técnica Bootstrap al considerarlas como independientes, los intervalos de confianza serán subestimados, saliendo menores de lo que son en realidad. Con Subset Bootstrap este error se corregirá, pues se considerarán subconjuntos de datos dependientes, y al tener menos información intrínseca, la incertidumbre será mayor.

5

Experimentos realizados y resultados

En este capítulo se hará un análisis de los resultados obtenidos a partir de las técnicas para hacer más robusto un modelo. En primer lugar se detallará la herramienta usada para las pruebas y el conjunto de datos utilizado, y a continuación se mostrará el análisis de resultados con las técnicas de la sección anterior.

5.1. Herramientas utilizadas

Para la elaboración de este TFG se ha utilizado **MATLAB** [25]. Matlab es una herramienta de cálculo, simulación y modelado matemático que ofrece un entorno de desarrollo integrado con lenguaje de programación propio. Este es capaz de trasladar cualquier problema a un punto de vista matemático. Además, es una herramienta muy potente para la visualización de resultados gráficos.

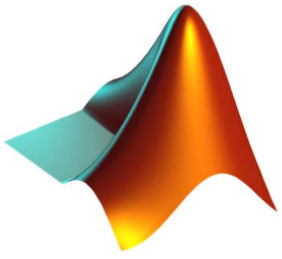

Figura 5.1: *Logotipo Matlab. [9]*

El uso de esta herramienta ha sido gracias a la Licencia Campus firmada entre la Universidad Autónoma de Madrid y la empresa MathWorks, que permite el uso de esta herramienta a cualquier miembro de la comunidad universitaria.

5.2. Conjunto de datos utilizado

El ejemplo con el que vamos a analizar los resultados es el análisis forense de vidrios.

Tendremos una base de datos con los LRs de pares de vidrios, distinguiendo los que provienen de la misma fuente y los que provienen de fuentes diferentes. Esta contará con una muestra de 62 vidrios diferentes y 11346 logLRs de las comparaciones entre todos esos vidrios, de los cuáles 3782 serán de la misma fuente y 7564 de fuentes diferentes. Estos vidrios provienen de una base de datos pública, que se puede encontrar en [26]. Además, el modelo utilzado para el cálculo está descrito como modelo MVK en [27].

El conjunto de datos además de indicarnos los vidrios comparados y el valor de la comparación, nos indicará si se trata de una comparación entre vidrios de la misma fuente o de fuentes distintas, es decir, tiene disponibles etiquetas en las que se sabe el valor real de la hipótesis en cada comparación, o etiquetas de "ground-truth".

En la práctica, tendremos un vector con todos los LRs mezclados *("logLRs")*, y otro vector *"keys"* con los valores de 0 y 1 en cada posición coincidente con el vector logLRs, de manera que si hay un 1, en esa misma posición de logLR se tratará de comparación entre vidrios de la misma fuente, y si, por el contrario el vector keys tiene un 0, en esa posición el valor de logLRs será entre vidrios de diferentes fuentes. Además, contaremos con un último vector llamado *"comparisons"* en el que, al igual que el anterior, cada posición del vector coincidirá con el LR de la posición del vector logLRs y con el 0 o 1 del vector keys. En el vector comparisons será donde indicaremos cuáles han sido los vidrios comparados en cada posición para obtener ese LR.

Antes de aplicar las técnicas de robustez, podemos representar la muestra de vidrios que tenemos en términos de curvas ECE, tal y como se muestra en la figura 5.2:

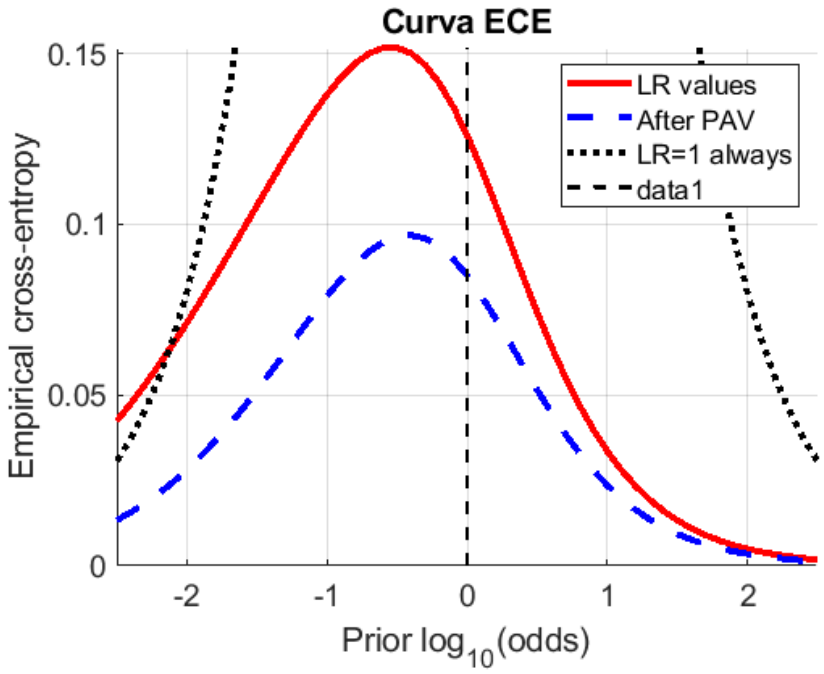

Figura 5.2: *Curva ECE para el conjunto de datos de los vidrios.*

Como vemos en la figura 5.2, se muestran todas las curvas estudiadas: la sólida que representa la entropía cruzada, la punteada o curva neutral, que tal y como vemos, se encuentra por encima de nuestra curva roja sólida para que tenga algo de discriminación. Y, por último, la curva azul discontinua extraída a partir del algoritmo PAV que, como vemos, aunque es bastante pequeña,

es decir, es mucho menor que la curva negra punteada que vale 1 siempre en x = 0, teniendo un poder de discriminación bastante alto, está bastante lejos de la línea roja, empeorando así la calibración.

5.3. Técnica Bootstrap

5.3.1. Pruebas básicas

Los primeros experimentos realizados han sido llevados a cabo para entender y mostrar bien los pasos para crear una muestra bootstrap. Por eso, hemos decidido explicarlo a partir de valores concretos de la curva ECE, no en toda la curva completa, en concreto del valor central o C_{llr} para la curva de la entropía cruzada y del $C_{llr}min$ o valor central de la curva óptima, también llamada bien calibrada (curva azul punteada). Para ello, seguiremos los pasos explicados en el apartado 4.2.1 de este TFG.

Entropía cruzada

1. En concreto para este ejemplo, como hemos dicho antes, el punto elegido ha sido el punto llamado C_{llr}, es decir, el punto medio en el que $P(\theta_p) = P(\theta_d) = 0{,}5$ para la curva roja, o también dicho, el punto en el que Prior log 10 (odds) sea 0.

   Así, en este primer paso y para nuestro ejemplo concreto, el valor C_{llr} calculado a partir del conjunto de datos dado ha sido un $C_{llr} = 0{,}12$. Este valor se puede ver aproximado en la figura 5.2.

2. En segundo lugar crearemos la muestra bootstrap a partir de nuestra muestra inicial.

   La muestra bootstrap se ha hecho de manera independiente para los LRs de la misma fuente y para los de fuentes diferentes utilizando la función de Matlab llamada *"datasample"*. Así, hemos cogido dos muestras del tamaño original de ambos. Es decir, de los 3782 valores de la misma fuente, hemos cogido 3782 valores con reemplazamiento, y de los 7564 valores de distintas fuentes, hemos cogido exactamente 7564 valores con reemplazamiento para nuestra primera muestra.

3. En tercer lugar, estimaremos el valor C_{llr} a partir de las muestras de fuentes iguales y fuentes diferentes obtenidas en el apartado anterior.

4. En cuarto y último lugar, repetiremos los pasos 2 y 3 varias veces. En este ejemplo en concreto se ha elegido un número bastante alto a repetir para que los intervalos sean lo máximo precisos posibles. El número de veces a repetir ha sido 100000.

Así, tal y como explicábamos en el apartado 4.2.2, con todos esos valores C_{llr} obtenidos, lo que haremos será ponerlos en orden. Para ver mejor este proceso, se ha creado un histograma que se muestra en la figura 5.3.

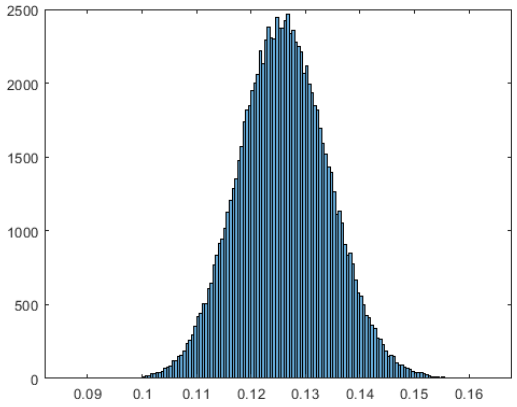

Figura 5.3: *Histograma de valores C_{llr} obtenidos a partir de una muestra bootstrap con 100000 repeticiones para un conjunto de datos de vidrios.*

Finalmente, podemos establecer los intervalos de confianza para el valor concreto C_{llr}. Puesto que en este ejemplo básico lo hemos hecho a partir de un solo valor escalar, los intervalos de confianza serán también valores escalares.

Para calcular el intervalo de confianza de por ejemplo el 90 %, a partir del histograma de la figura 5.3 cogeremos el 5 % por abajo, siendo éste el percentil del 5 % y cogeremos el 5 % por arriba, siendo éste el percentil del 95 % y así, obtendremos los dos valores de nuestro intervalo de confianza en el punto medio de la gráfica, es decir, cuando $P(\theta_p) = P(\theta_d) = 0{,}5$.

En la figura 5.4 se muestran los puntos de los intervalos de confianza para el valor sacado de la muestra inicial $C_{llr} = 0{,}12$, es decir, se muestran los dos puntos entre los que se se estima que estará nuestro valor al variar la calidad o cantidad de datos.

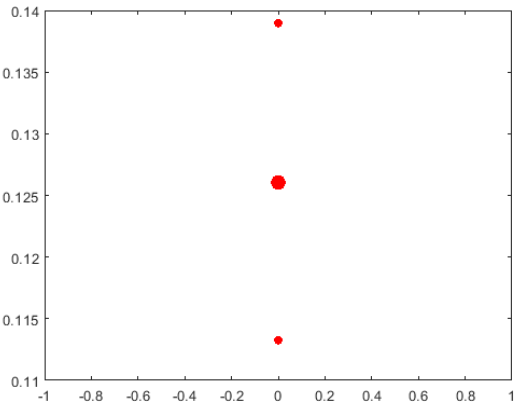

Figura 5.4: *Gráfica de puntos que muestra el valor escalar C_{llr} y sus correspondientes intervalos de confianza del 90 % en el mismo eje de abscisas (X) para 100000 repeticiones.*

En caso de coger una muestra bootstrap con una repetición menor, es decir, con menos iteraciones, el intervalo podrá cambiar cada vez que se ejecute el programa, por lo que no será del todo robusto. Es por eso que es importante, repetirlo una cantidad de veces elevado para evitar desajustes en el intervalo de confianza. En el Anexo A, mostraremos unos ejemplos con 10 y 1000 repeticiones para ver cómo cambia el intervalo de confianza. Si nos damos cuenta, parece que los pares de valores entre los que se estima que estará nuestro valor convergen a $C_{llr} = 0{,}14$

por arriba y a $C_{llr} = 0{,}11$ por abajo, por lo que tendremos unos intervalos de confianza fiables.

Entropía óptima

A continuación, repetiremos los pasos anteriormente explicados pero con una diferencia importante, los intervalos de confianza a sacar, no van a ser intervalos sobre la curva de la entropía cruzada, sino sobre la curva óptima sacada a partir del algoritmo PAV. En concreto sobre el valor central de esta curva, llamado $C_{llr}min$.

En el primer paso del proceso, tal y como podemos ver en la figura 5.2, el $C_{llr}min$ estará en $C_{llr}min = 0{,}085$. Después de ejecutar los pasos 2 y 3 un número elevado de veces, podemos sacar, al igual que antes, los intervalos de confianza del 90 %. En la figura 5.5 mostraremos el histograma de todos los valores de $C_{llr}min$ calculados en cada iteración, y en la figura 5.6 el intervalo de confianza del 90 % de acierto para ese valor.

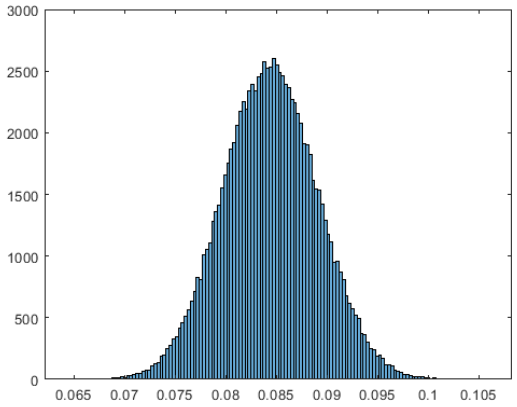

Figura 5.5: *Histograma de valores $C_{llr}min$ obtenidos a partir de una muestra bootstrap para un conjunto de datos de vidrios.*

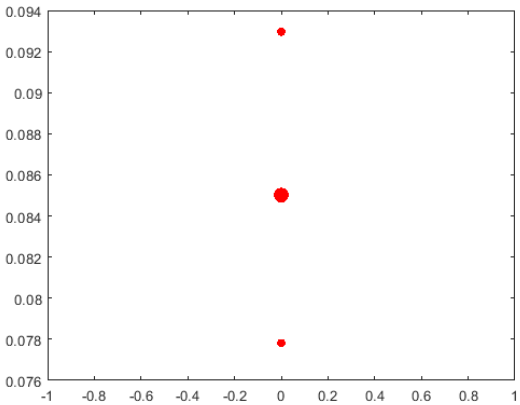

Figura 5.6: *Gráfica de puntos que muestra el valor escalar $C_{llr}min$ y sus correspondientes intervalos de confianza del 90 % en el mismo eje de abscisas (X).*

5.3.2. Pruebas avanzadas: Bootstrap con todos los puntos

En el ejemplo anterior hemos seguido paso a paso el proceso para construir una muestra bootstrap para un valor escalar. En este apartado a partir de la misma base de datos de los vidrios haremos una muestra bootstrap para todos los valores de la curva, para poder así, construir unos intervalos de confianza que serán curvas también. Al igual que antes, calcularemos los intervalos de confianza para la curva de entropía cruzada(roja) y para la curva óptima(azul).

Entropía cruzada

El proceso a seguir ha sido el mismo que en el apartado anterior:

1. En primer lugar estimaremos la curva de entropía cruzada según la ecuación 3.5 con todos los valores de la base de datos, distinguiendo entre los de la misma fuente(targetScores) y los de fuentes diferentes(nonTargetScores). La curva resultado, será la que se muestra en la figura 5.7.

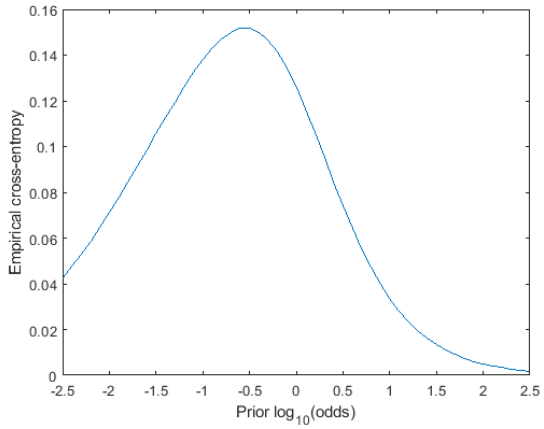

Figura 5.7: *Curva ECE de la entropía cruzada como la función logaritmica de las probabilidades a priori para el conjunto de datos de los vidrios.*

2. En segundo lugar crearemos la muestra bootstrap a partir de nuestra muestra inicial. Igual que antes, el procedimiento será coger una muestra con la misma cantidad de LRs que la muestra inicial, es decir, 3782 LRs de la misma fuente y 7564 LRs de fuentes distintas.

3. En tercer lugar, estimaremos la curva de entropía cruzada para la muestra obtenida en el paso anterior.

4. En cuarto y último lugar, repetiremos los pasos 2 y 3 varias veces para conseguir varias estimaciones de la curva de la entropía cruzada. En este ejemplo el número de estimaciones han sido 1000.

Al estar estimando en cada iteración todos los valores que forman una curva, tendríamos histogramas para todos los valores que forman la curva como en la figura 5.3. Al hacer el percentil 5 % y 95 % de cada uno de ellos obtenemos, como antes, el intervalo de confianza del 90 % de acierto, es decir las dos curvas entre las que se estima que estará nuestra curva de entropía cruzada al variar el conjunto de datos con un acierto del 90 %. Esto se muestra en la figura 5.8:

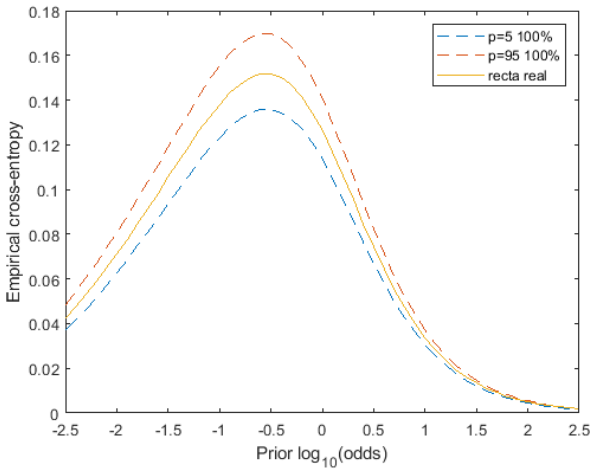

Figura 5.8: *Curva entropía cruzada y sus correspondientes intervalos de confianza del 90 % de acierto con Bootstrap para 1000 repeticiones y cogiendo el 100 % del número de la muestra inicial.*

A continuación, veremos cómo varían los intervalos de confianza si en vez de coger en cada iteración una muestra del tamaño de los datos iniciales, cambiamos este tamaño.

En la figura 5.9 podemos ver la variación del intervalo de confianza, cogiendo el 10 % del tamaño de los datos, el 50 %, el 90 % y el 100 % (como antes).

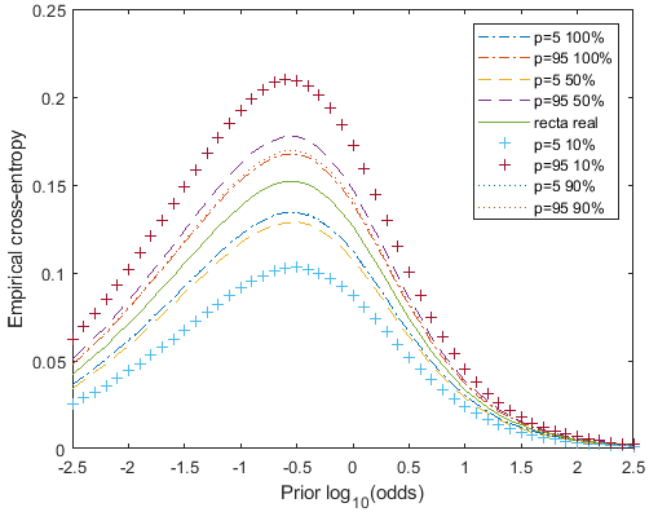

Figura 5.9: *Curva entropía cruzada y sus correspondientes intervalos de confianza del 90 % de acierto con Bootstrap para 1000 repeticiones y cogiendo el 100 %, 90 %, 50 % y 10 % del número de la muestra inicial.*

Tal y como vemos en la imagen anterior, a medida que aumenta el tamaño de la muestra bootstrap, menor es el intervalo de confianza. Esto tiene sentido pues, por lo general, mientras más observaciones haya, más estrecho será el intervalo alrededor del estadístico de la muestra, reduciendo así el margen de error. Pues, si el intervalo de confianza es demasiado ancho, no se puede estar muy seguro del valor real de un parámetro.

En el Anexo A mostraremos una gráfica completa con los intervalos de confianza variando la cantidad de los datos: desde el 10 % al 90 % de 10 en 10 de los datos de la muestra inicial.

Entropía óptima

El procedimiento para hallar los intervalos de confianza será exactamente el mismo que en el apartado anterior, pero se hará sobre los valores de la curva óptima obtenida a partir del algoritmo PAV.

Así, en el primer paso del proceso se calcula la curva óptima a partir de los valores de la muestra inicial de los vidrios. Esta curva se muestra en la figura 5.10.

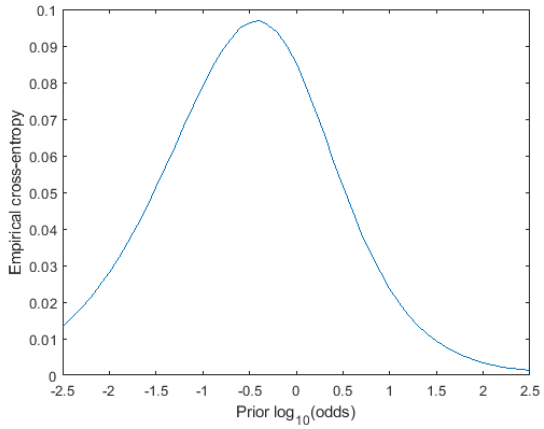

Figura 5.10: *Curva ECE de la entropía óptima como la función logaritmica de las probabilidades a priori para el conjunto de datos de los vidrios.*

Finalmente, después de coger diferentes muestras, calcularemos sus intervalos de confianza. Dependiendo del tamaño de las muestras obtenidas, los intervalos serán mejores o peores, como hemos visto en el apartado anterior. A continuación se muestran en la figura 5.11 los intervalos de confianza para una muestra cogiendo un tamaño igual al 100 % inicial, y en la figura 5.12 como varían estos intervalos al coger un tamaño del 10 %, 50 %, 90 % y 100 %.

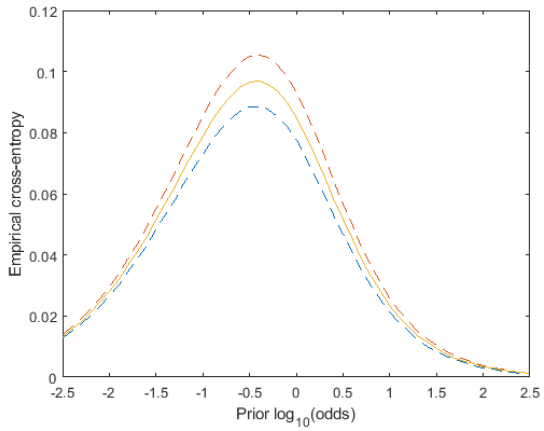

Figura 5.11: *Curva entropía óptima y sus correspondientes intervalos de confianza del 90 % de acierto con Bootstrap para 1000 repeticiones y cogiendo el 100 % del número de la muestra inicial.*

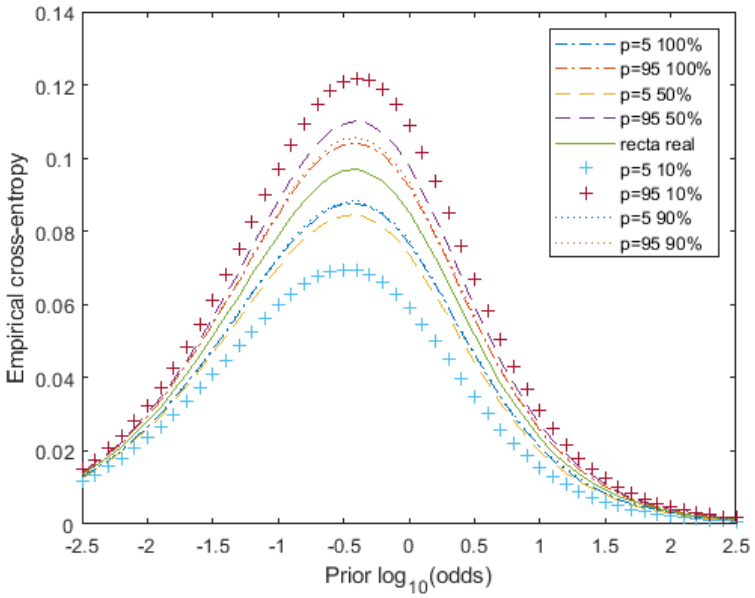

Figura 5.12: *Curva entropía óptima y sus correspondientes intervalos de confianza del 90 % de acierto con Bootstrap para 1000 repeticiones y cogiendo el 100 %, 90 %, 50 % y 10 % del número de la muestra inicial.*

Uniendo las dos curvas

En las figuras 5.13 y 5.14 podemos ver un simple ejemplo en el que mostraremos las dos curvas anteriormente explicadas con un determinado intervalo de confianza del 90 % de acierto cogiendo una cantidad de datos igual a la muestra inicial. En realidad es el resultado de juntar las figuras 5.8 y 5.11.

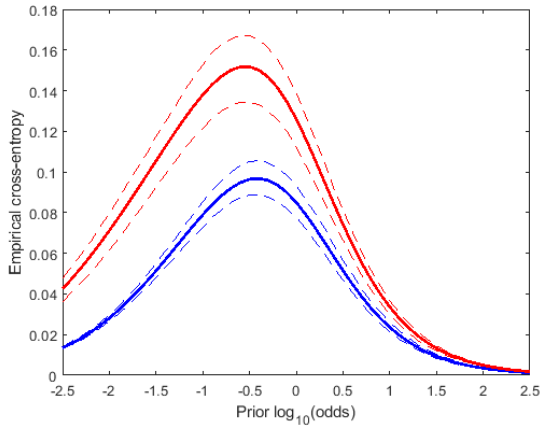

Figura 5.13: *Curva entropía óptima y entropía cruzada con sus correspondientes intervalos de confianza del 90 % de acierto con Bootstrap cogiendo el 100 % del número de la muestra inicial.*

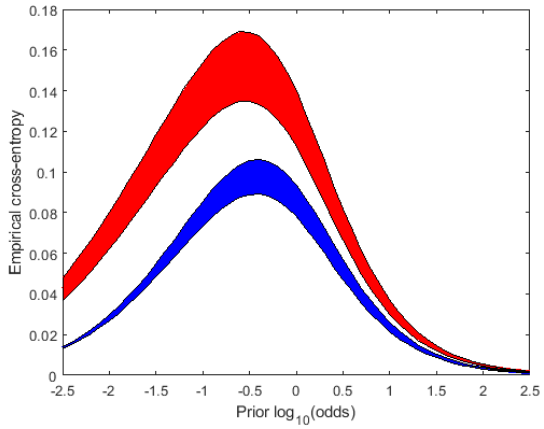

Figura 5.14: *Relleno intervalos de confianza del 90 % de acierto con Bootstrap para la curva de entropía óptima y de entropía cruzada cogiendo el 100 % del número de la muestra inicial.*

5.4. Técnica Subset Bootstrap

5.4.1. Modificación del dataset

Tal y como hemos mencionado en el apartado 4 de este TFG, la técnica bootstrap solo se puede hacer con datos independientes. Así, en los experimentos anteriores hemos supuesto que todos los datos eran independientes entre sí, pero en realidad no lo son. Por ejemplo, el LR calculado por los vidrios 1 y 3, y el LR calculado por los vidrios 1 y 2, serán dependientes entre sí, pues ambos dependen del vidrio 1.

En nuestro dataset tenemos LRs comparando todos con todos los vidrios. Es por eso que habrá que modificarlo para conseguir el Subset Boostrap.

Puesto que tenemos 62 vidrios, haremos subconjuntos de 5, dejando fuera 2 vidrios para poder tener un número múltiplo de 5. Así, tendremos un total de 12 subconjuntos.

Sabemos que el Subset Bootstrap consiste en hacer subconjuntos de muestras en los que dentro de cada subconjunto las muestras son dependientes entre sí, pero todas ellas son independientes con las de otros subconjuntos. Teniendo en cuenta como funciona el Subset Bootstrap, en cada subconjunto tendremos las comparaciones entre esos 5 vidrios del subconjunto y entre sí mismos. Finalmente, nos damos cuenta que hay que descartar un número considerable de comparaciones de cada vidrio con el resto de vidrios que no pertenecen a su subconjunto.

Podríamos entenderlo mejor con la imagen 5.15 en la que cada cuadrado negro constituye un subconjunto, y en cada subconjunto están los LRs entre los vidrios de ese subconjunto. El resto, lo blanco que forma la figura, serán los LRs entre vidrios de distintos subconjuntos, por lo que nos desharemos de ellos:

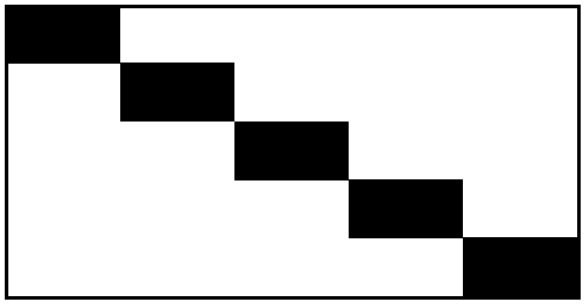

Figura 5.15: *Ejemplo visual del Subset Bootstrap.*

Así, eliminando los LRs entre vidrios de distintos subconjuntos, mostraremos como varía el Bootstrap normal. Aún no haremos el subset bootstrap, simplemente mostraremos como varían los intervalos de confianza en el bootstrap cogiendo como muestra inicial, en vez de todos los LRs, los LRs que pertenecen a algún subconjunto. En las figuras 5.16 y 5.17 se mostrarán los intervalos de confianza cogiendo el 10 %, 50 % ,90 % y 100 % del tamaño modificado, es decir, el correspondiente a los cuadrados negros de la imagen 5.15 de 12 subconjuntos.

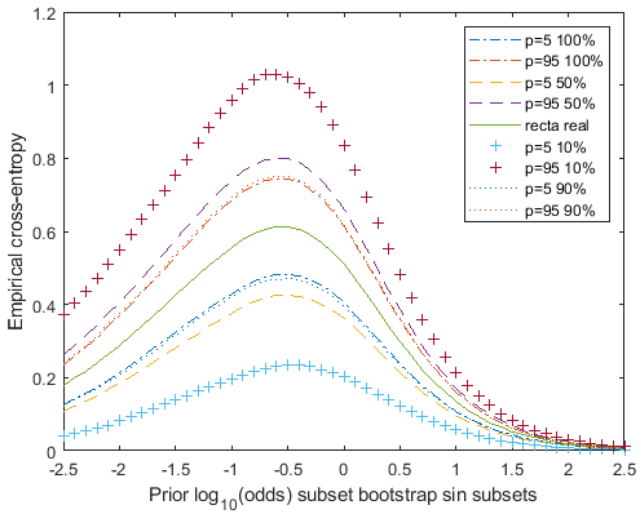

Figura 5.16: *Curva entropía cruzada y sus correspondientes intervalos de confianza del 90 % de acierto con Bootstrap para 1000 repeticiones y cogiendo el 100 %, 90 %, 50 % y 10 % de la muestra modificada.*

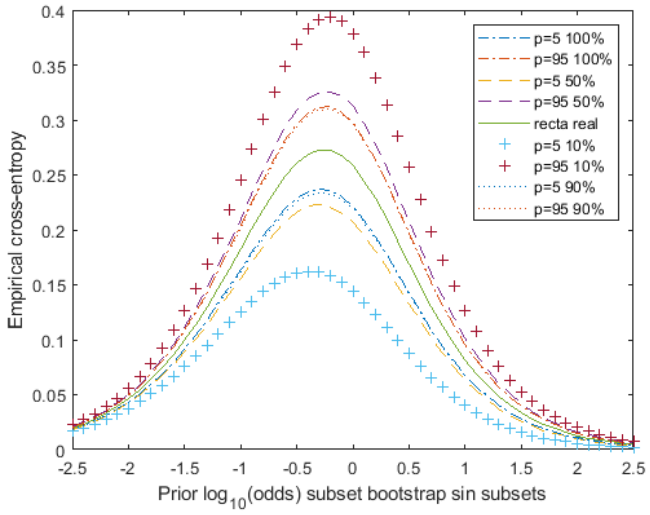

Figura 5.17: *Curva entropía óptima y sus correspondientes intervalos de confianza del 90 % de acierto con Bootstrap para 1000 repeticiones y cogiendo el 100 %, 90 %, 50 % y 10 % de la muestra modificada.*

Como vemos en las figuras 5.16 y 5.17, los intervalos de confianza aumentan con respecto a las figuras 5.9 y 5.12, esto es debido a la muestra de la que se parte para coger las muestras bootstrap. En las figuras 5.16 y 5.17 es normal que los intervalos de confianza salgan mayores, puesto que se ha disminuido el número de datos, y aunque cojamos el 100 % del tamaño de la muestra, esta muestra modificada parte de menos datos que la muestra inicial, aumentando así el margen de error.

5.4.2. Subset bootstrap con la muestra modificada

En este apartado a partir de la muestra de datos de los vidrios modificada en el apartado anterior haremos una muestra subset bootstrap para todos los valores de la curva, para poder así, construir unos intervalos de confianza y medir la robustez del modelo. Esta técnica la aplicaremos, igual que antes, tanto a la curva roja o curva de entropía cruzada como a la curva azul o curva óptima obtenida a partir del algoritmo PAV.

El proceso a seguir ha sido el explicado en el 4.3 de este TFG:

1. Se calcularán las curvas de entropía cruzada y entropía óptima a partir de los nuevos datos que tenemos. Estas curvas las podemos ver en las figuras 5.16 y 5.17 del apartado anterior.
2. Se creará la muestra bootstrap con reemplazamiento de cada subconjunto. En este ejemplo cogeremos una cantidad del 100 % del número de datos inicial, es decir, puesto tendremos un total de 12 subconjuntos cogeremos 12 de esos subconjuntos con reemplazamiento.
3. Se estimarán de nuevo las dos curvas a partir de la muestra extraída en el apartado anterior.
4. Se repetirán los pasos 2 y 3. En este caso el número de veces a repetir ha sido 1000.

A continuación, mostraremos los intervalos de confianza del 100 % de los datos aplicando la técnica subset bootstrap a nuestro dataset de vidrios.

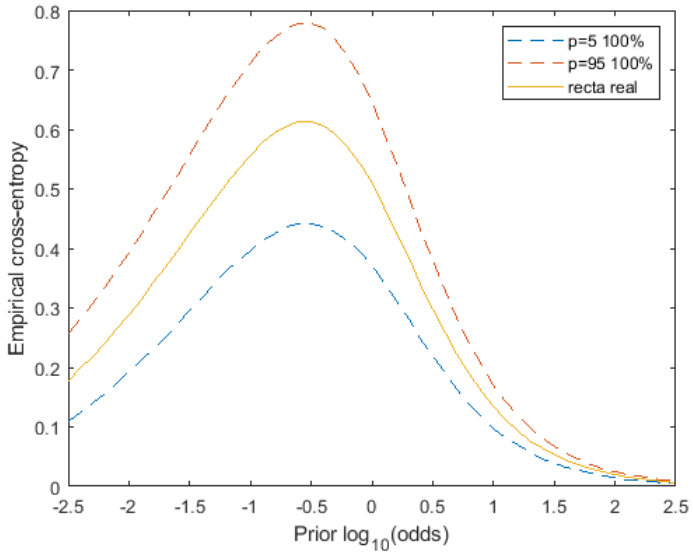

Figura 5.18: *Curva entropía cruzada y sus correspondientes intervalos de confianza del 90 % de acierto con Subset Bootstrap para 1000 repeticiones y cogiendo el 100 % de los subconjuntos.*

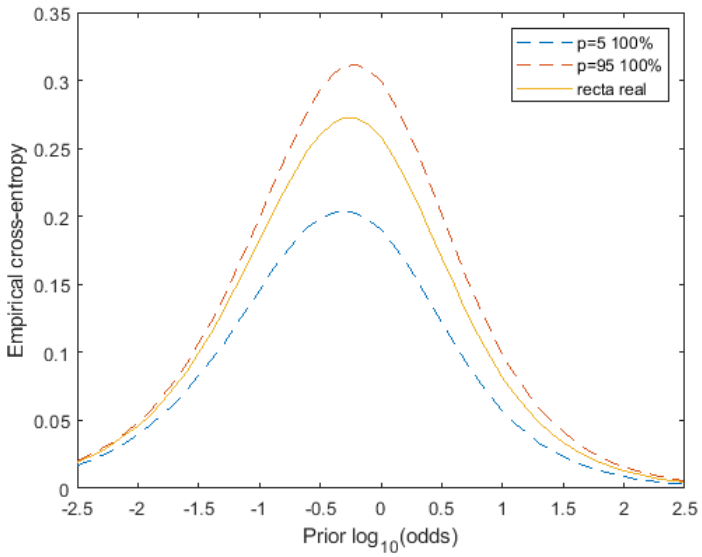

Figura 5.19: *Curva entropía óptima y sus correspondientes intervalos de confianza del 90 % de acierto con Subset Bootstrap para 1000 repeticiones y cogiendo el 100 % de los subconjuntos.*

Si comparamos la técnica Bootstrap y Subset Bootstrap a partir de las figuras 5.8 y 5.18 para la entropía cruzada y 5.11 y 5.19 para la entropía óptima nos damos cuenta de que los intervalos de confianza varían bastante para un mismo conjunto de datos.

Cogiendo un tamaño igual al de la muestra inicial (en el primer caso cogiendo 11346 LRs con repetición en cada iteración, y en el segundo caso cogiendo 12 subconjuntos con repetición) los intervalos de confianza de las figuras en las que se ha aplicado bootstrap son menores que en los que se ha aplicado subset bootstrap. Esto quiere decir que, en efecto, los LRs son dependientes.

Hemos considerado como independientes todos los LRs para utilizar la técnica Bootstrapping, pero hemos subestimado los intervalos, creando unos intervalos más pequeños de lo que

realmente son. Ahora nos damos cuenta de que son dependientes, puesto que con ésta técnica tenemos una mayor varianza y por lo tanto, estaría mal usar la técnica Bootstrapping para evaluar la robustez del método. La técnica adecuada sería el Subset Bootstrap.

A continuación, mostraremos los intervalos de confianza del subset bootstrap variando la cantidad de datos cogidos en la muestra, en concreto cogiendo el 10 % de los subconjuntos (que en este caso es 1), el 50 %, el 90 % y el 100 % (como antes).

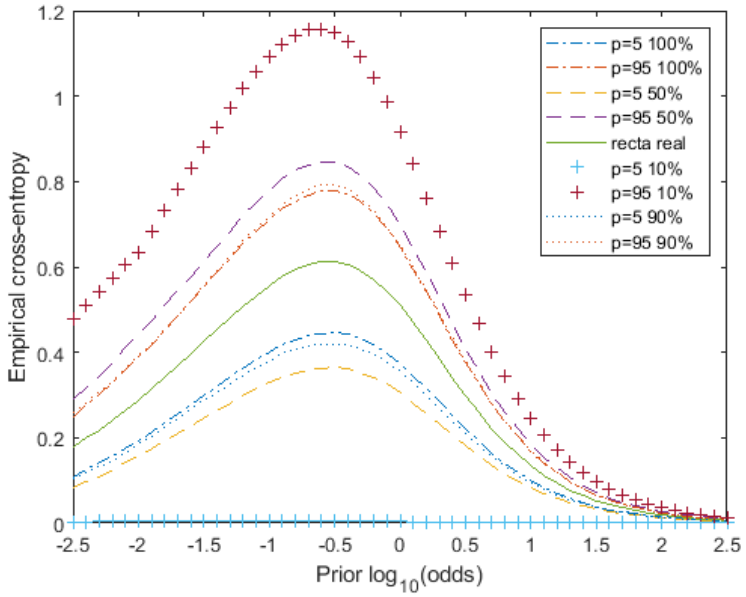

Figura 5.20: *Curva entropía cruzada y sus correspondientes intervalos de confianza del 90 % de acierto con Subset Bootstrap para 1000 repeticiones y cogiendo el 100 %, 90 %, 50 % y 10 % de los subconjuntos.*

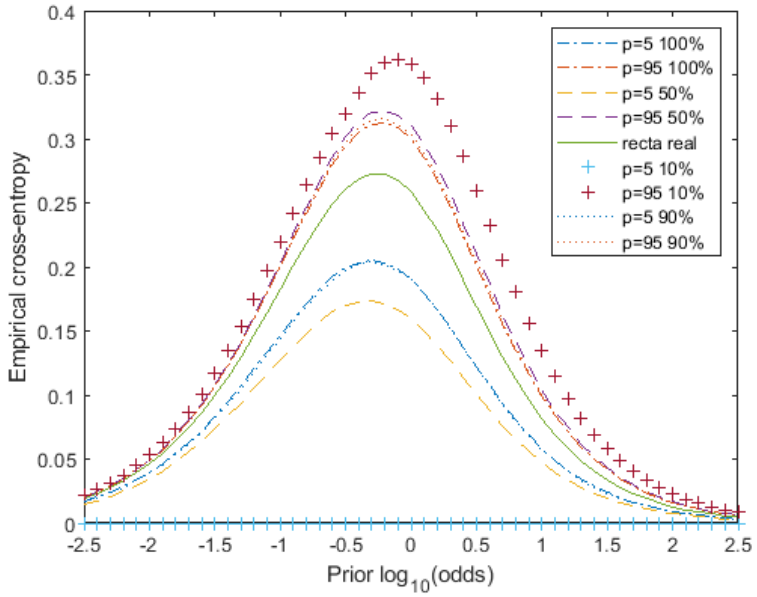

Figura 5.21: *Curva entropía óptima y sus correspondientes intervalos de confianza del 90 % de acierto con Subset Bootstrap para 1000 repeticiones y cogiendo el 100 %, 90 %, 50 % y 10 % de los subconjuntos.*

Tal y como vemos en las imágenes anteriores, al igual que pasaba con la técnica Bootstrap, a medida que aumenta el tamaño de la muestra subset bootstrap en cada iteración menor es el intervalo de confianza.

En el Anexo B mostraremos las gráficas completas con los intervalos de confianza variando la cantidad de los datos: desde el 10 % al 90 % de 10 en 10 de los subconjuntos.

Finalmente, mostraremos, al igual que en Bootstrap, la unión de las dos curvas aplicando el método Subset bootstrap con un intervalo de confianza del 90 % de acierto cogiendo los 10 subconjuntos.

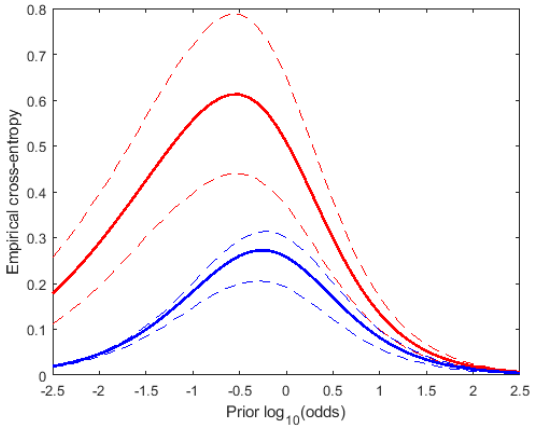

Figura 5.22: *Curva entropía óptima y entropía cruzada con sus correspondientes intervalos de confianza del 90 % de acierto con Subset Bootstrap cogiendo el 100 % del número de la muestra inicial.*

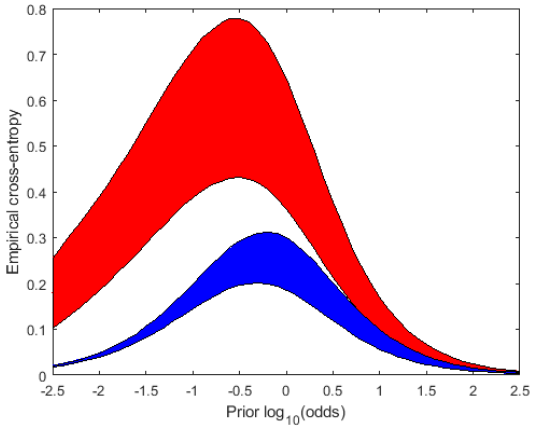

Figura 5.23: *Relleno intervalos de confianza del 90 % de acierto con Subset Bootstrap para la curva de entropía óptima y de entropía cruzada cogiendo el 100 % del número de la muestra inicial.*

6

Conclusiones y Trabajo futuro

6.1. Conclusiones

En todo juicio en el que entra en juego la ciencia forense hay un proceso de análisis de evidencias. Este proceso es el que desde la comparación de dos muestras, se efectúa un informe aportando información sobre si pertenecen a la misma fuente o no. La metodología de decisión bayesiana aporta un marco formal, mediante el cual la prueba forense se valora mediante un ratio de verosimilitudes, o LR. Además del desarrollo del método, es importante su validación, que se divide en dos partes: (1) procesos relacionados con la selección del modelo y la fase de entrenamiento, y (2) validación utilizando otro conjunto de datos. Mientras que en la primera parte se evalúan los rendimientos primarios como calibración y poder discriminativo, en la segunda, se evaluarán además, los rendimientos secundarios como coherencia, generalización y robustez. En el presente TFG nos hemos centrado en la robustez.

Sabiendo que la robustez se refiere en concreto a la estabilidad del método al variar algunas condiciones como calidad o cantidad de muestras en la base de datos, una manera de medir esta precisión y robustez ha sido con intervalos de confianza. Para encontrar estos intervalos, en este TFG hemos propuesto usar las técnicas bootstrap y subset bootstrap. Hemos comparado las dos técnicas y las ventajas de cada una.

Como hemos visto, la técnica bootstrap solo puede ser utilizada cuando los datos son independientes entre sí. Si tenemos muestras con dependencia, tal y como sucedía con las muestras de nuestro ejemplo de vidrios, no podremos utilizar bootstrap normal y es cuando introducimos el subset bootstrap. Así, dado nuestro dataset de vidrios, modificaremos la muestra para poder crear subconjuntos de LRs de vidrios dependientes entre sí dentro del mismo subconjunto e independientes al resto de LRs de otros subconjuntos. Con el dataset modificado se podrá utilizar la técnica subset bootstrap.

Al comparar los márgenes de error entre las dos técnicas descritas en concreto en nuestro dataset, nos damos cuenta de que los intervalos de confianza del bootstrap son menores. Como hemos visto bootstrap suele subestimar los intervalos de confianza, creando unos intervalos menores de lo que realmente son. La razón de esto es que un conjunto de LRs que son dependientes entre sí nos aporta menos información que otros totalmente independientes, y esto provocará una mayor varianza, como se ha mostrado en el Trabajo.

Hemos construido los intervalos de confianza con los LRs de fuentes iguales y de fuentes diferentes y los hemos mostrado sobre una curva ECE, que es una medida de rendimiento del modelo utilizado. Esta curva nos indica además, el buen poder de discriminación y la buena calibración que tiene nuestro modelo. Por lo tanto, podemos decir que lo que se pretende en este TFG es dar una medida de robustez sobre una medida de rendimiento del marco bayesiano de interpretación para una muestra concreta, en este caso una muestra de vidrios.

6.2. Trabajo futuro

A partir de este trabajo surgen nuevas líneas de investigación:

- **Nuevas bases de datos**: Como hemos podido ver este trabajo solo se ha investigado y probado con una base de datos de 62 vidrios. Por eso, se propone extender y conocer resultados al probar estas técnicas con nuevas bases de datos, incluso más complejas y de mayor cantidad.
- **Utilizar más factores de variabilidad**: Ya hemos visto que para bases de datos de vidrios cuando son éstos dependientes, el subset bootstrap consigue buenos márgenes de error en los que podemos confiar. El objetivo por lo tanto sería probar a comparar vidrios en los que no dependiera solo del número que se compara(vidrio1 con el vidrio3), sino de otros factores como del tipo de procedencia por ejemplo, es decir, en una base de datos distinguir entre los vidrios provenientes de vasos, o los que provienen de ventanas.
- **Probar con otros modelos**: Existen muchos otros, que dependiendo la base de datos que usemos serán mejores o peores y nos darán distintos intervalos de confianza, además de un poder discriminativo distinto y de distinta calibración. El objetivo futuro será, por lo tanto, comparar los resultados con otros modelos y estudiar el cambio en los intervalos de confianza.
- **Utilizar otras medidas de rendimiento**: Por último, proponemos extender la investigación y probar las técnicas explicadas en otras medidas de rendimiento distintas. En este trabajo, solo se ha mostrado la eficacia de estas técnicas sobre las curvas ECE, por lo que sería conveniente buscar y analizar sobre qué otras medidas se podrían probar. Algunos ejemplos son el área bajo la curva ROC como medida de discriminación, o las curvas Tippett. [14]

Bibliografía

[1] D. Ramos. *Forensic evaluation of the evidence using automatic speaker recognition systems.* PhD thesis, Depto. de Ingenieria Informatica, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain, 2007. Available at http://atvs.ii.uam.es.

[2] D. Ramos. Valoración de evidencias forenses ii. tema 3: Validación de métodos de cálculo de lr. *Asignatura EPS-UAM*, 2015.

[3] Juan Maroñas Molano. *Comparación de algoritmos de cálculo de ratios de verosimilitudes para interpretación forense.* PhD thesis, Depto. de Ingenieria Informatica, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain, 2015.

[4] Daniel Ramos, Javier Franco-Pedroso, Alicia Lozano-Diez, and Joaquin Gonzalez-Rodriguez. Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy*, 2018. To appear.

[5] Daniel Ramos and Joaquin Gonzalez-Rodriguez. Reliable support: measuring calibration of likelihood ratios. *Forensic Science International*, 230:156–169, 2013.

[6] D. Ramos. Tema 2: Tratamiento estadístico de errores. *Asignatura de "Instrumentación Electrónica"de la antigua Ingeniería de Teleco de la EPS-UAM.*, 2013.

[7] G. Martinez. Conjunto de clasificadores. *Asignatura de "Fundamentos de Aprendizaje Automático"de Ingeniería Informática EPS-UAM.*

[8] Explicación de la función normal y algunos ejemplos. `https://matematicasconmuchotruco.wordpress.com/2014/07/`.

[9] Logotipo de matlab. `https://ccm.net/faq/2762-concatenate-vectors-or-matrices-under-matlab`.

[10] P. Hart R. Duda and D. Stork. Pattern classification. *John Wiley and Sons*, (2), 2000.

[11] Javier Franco-Pedroso, Daniel Ramos, and Joaquin Gonzalez-Rodriguez. Gaussian mixture models of between-source variation for likelihood ratio computation from multivariate data. *PLoS ONE*, 11(2):1–25, February 2016.

[12] S. Theodoridis and K. Koutroumbas. Pattern recognition. *Academic Press.*, (2), 2003.

[13] D. Ramos. Reliable support: Measuring calibration of likelihood ratios. *ATVS. Biometric Recognition Group Research Institute of Forensic Science and Security UAM.*, 2012.

[14] Didier Meuwly, Daniel Ramos, and Rudolf Haraksim. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276:142–153, July 2017.

[15] D. Ramos. Evaluation of likelihood ratios based on information theory, 2007. Invited presentation at One Day One Topic Seminar and Workshop - Forensic Evidence Evaluation. Institute of Forensic Research, Cracow, Poland. 22nd-23rd June 2007. http://www.ies.krakow.pl/oos2007/index.php?link=organisers.php.

[16] Explicación de la calibración y discriminación en regresión logística. `http://www.hrc.es/bioest/Reglog_10.html`.

[17] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia. Likelihood ratio calibration in transparent and testable forensic speaker recognition. In *Proc. of Odyssey*, 2006.

[18] David J.C MacKay. Information theory, inference, and learning algorithms. *Cambridge University Press.*, pages 137–177, 2003.

[19] D. van Leeuwen and N. Brümmer. An introduction to application-independent evaluation of speaker recognition systems. In Christian Müller, editor, *Speaker Classification*, volume 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*. Springer, Heidelberg - Berlin - New York, 2007.

[20] N. Brümmer. *Measuring, refining and calibrating speaker and language information extracted from speech.* PhD thesis, School of Electrical Engineering, University of Stellenbosch, Stellenbosch, South Africa, 2010. Available at http://sites.google.com/site/nikobrummer/.

[21] R. M. Bolle, N. K. Ratha, and S. Pankanti. Error analysis of pattern recognition systems–the subsets bootstrap. *Computer Vision and Image Understanding*, 93:1–33, 2004.

[22] Intervalos de confianza. `https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/basics/what-is-a-confidence-interval/`.

[23] Introducción a la creación de intervalos de confianza a partir de la técnica bootstrap. `https://anestesiar.org/2015/una-tarea-imposible-la-tecnica-de-bootstrapping/`.

[24] Distribuciones de muestreo. `https://wwwyyy.files.wordpress.com/2013/03/distribuciones-en-el-muestreo.pdf`.

[25] Descripción de matlab. `http://nereida.deioc.ull.es/~pcgull/ihiu01/cdrom/matlab/contenido/node2.html`.

[26] Base de datos pública de vidrios con la que hemos realizado los experimentos. `https://onlinelibrary.wiley.com/doi/book/10.1002/9781118763155`.

[27] Descripción del modelo utilzado para el cálculo de lrs en nuestro conjunto experimental. `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0149958`.

Glosario de acrónimos

- **ECE**: Curva de entropía cruzada. (Empirical Cross Entropy)
- **LR**: Razon de verosimilitudes. (Likelihood ratio)
- **FA**: Falso positivo. (False Accept)
- **FR**: Falso negativo. (False Reject)
- $\mathbf{C_{fa}}$: Coste del falso positivo. (False Accept Cost)
- $\mathbf{C_{fr}}$: Coste del falso negativo. (False Reject Cost)
- **PAV**: Pool Adjacent Violators.
- **SPSR**: Strictly Proper Scoring Rules.
- $\mathbf{C_{llr}}$: valor de ECE donde las probabilidades a priori se igualan.
- $\mathbf{C_{llr}^{min}}$: C_{llr} debido a la falta de discriminación.
- $\mathbf{C_{llr}^{cal}}$: C_{llr} debido a la falta de calibración.
- $\mathbf{ECE^{min}}$: Información perdida por la falta de discriminación.
- $\mathbf{ECE^{cal}}$: Información perdida por la falta de calibración.
- **C**: Media del coste debido a la falta de exactitud.
- $\mathbf{C^{cal}}$: Coste por la falta de calibración.
- $\mathbf{C^{dic}}$: Coste por la falta de discriminación.

Anexos

A

Anexo A: Bootstrap

A.1. Ejemplo bootstrap para un único valor de la curva ECE

En este apartado se muestran los resultados obtenidos variando el número de repeticiones para una muestra bootstrap del conjunto de datos de los vidrios. En concreto, los ejemplos se muestran para un único valor de la curva ECE. El elegido ha sido el valor medio, es decir, el C_{llr}.

A.1.1. Ejemplo con 10 repeticiones

A continuación se muestra el histograma de los valores de C_{llr} calculados con una muestra bootstrap de 10 repeticiones:

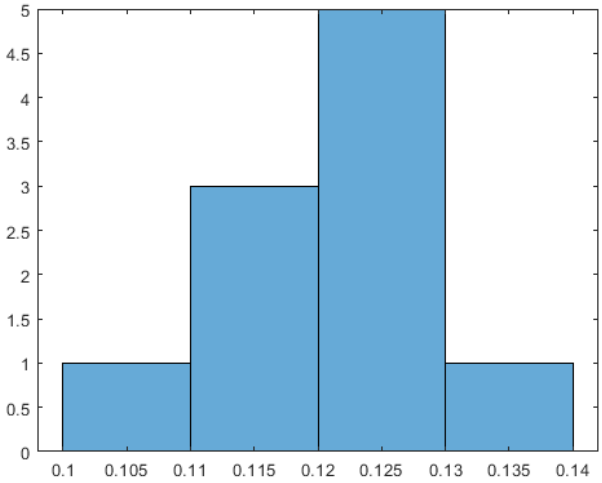

Figura A.1: *Histograma de valores C_{llr} obtenidos a partir de una muestra bootstrap con 10 repeticiones para un conjunto de datos de vidrios.*

En la siguiente figura, se mostrará el intervalo de confianza del 90 % a partir del histograma anterior. Puesto que son valores escalares, los intervalos de confianza serán simplemente los puntos de los percentiles 5 y 95.

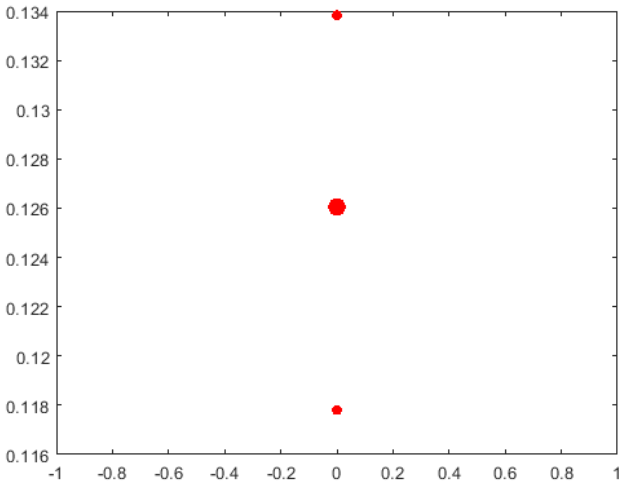

Figura A.2: *Gráfica de puntos que muestra el valor escalar c_{llr} y sus correspondientes intervalos de confianza del 90 % en el mismo eje de abscisas (X) para 10 repeticiones.*

A.1.2. Ejemplo con 1000 repeticiones

A continuación se muestra el histograma de los valores de C_{llr} calculados con una muestra bootstrap de 1000 repeticiones:

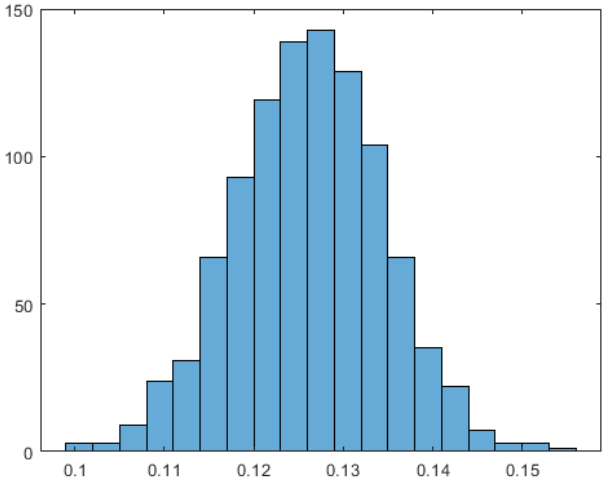

Figura A.3: *Histograma de valores C_{llr} obtenidos a partir de una muestra bootstrap con 1000 repeticiones para un conjunto de datos de vidrios.*

En la siguiente figura, se mostrará el intervalo de confianza del 90 % a partir del histograma anterior. Puesto que son valores escalares, los intervalos de confianza serán simplemente los puntos de los percentiles 5 y 95.

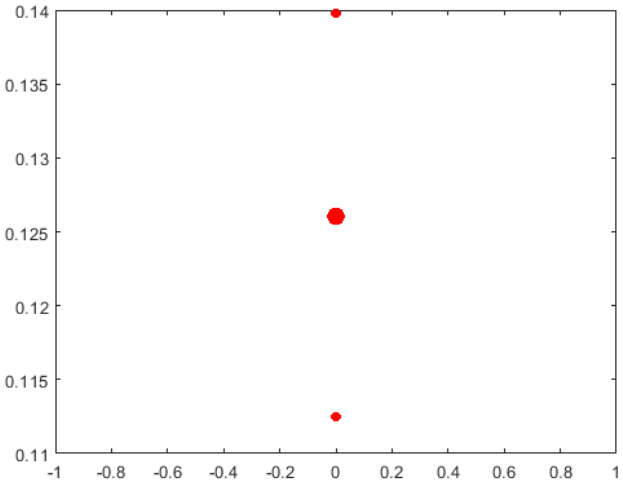

Figura A.4: *Gráfica de puntos que muestra el valor escalar c_{llr} y sus correspondientes intervalos de confianza del 90 % en el mismo eje de abscisas (X) para 1000 repeticiones.*

A.2. Ejemplo bootstrap de todos los valores de una curva

En este apartado se muestran los resultados obtenidos variando la cantidad de la muestra bootstrap. En las siguientes figuras mostraremos los intervalos de confianza cogiendo desde el 10 % de los datos hasta el 90 % y veremos como a medida que aumenta este porcentaje se va reduciendo el intervalo:

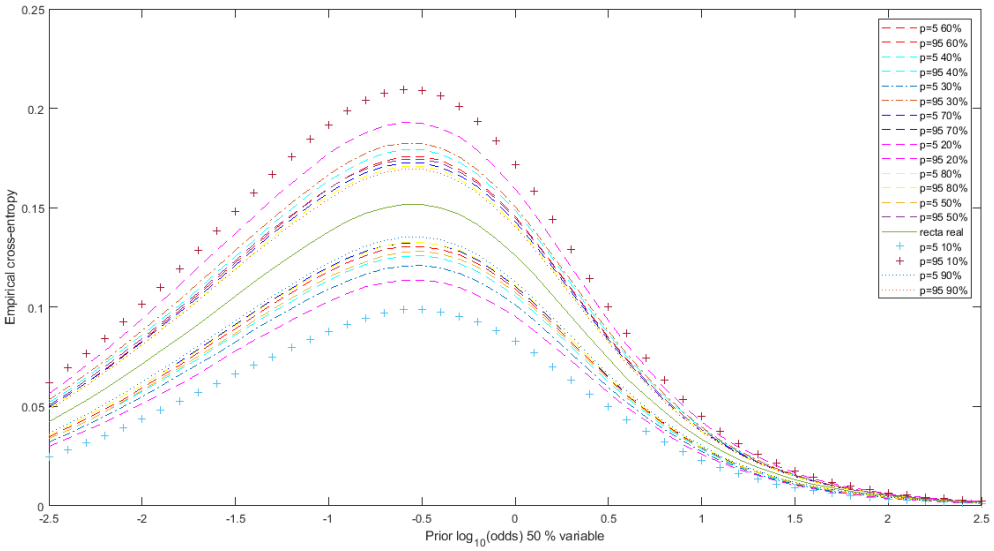

Figura A.5: *Curva entropía cruzada y sus correspondientes intervalos de confianza del 90 % de acierto con bootstrap para 1000 repeticiones y cogiendo desde el 90 % al 10 % de 10 en 10 de la muestra inicial.*

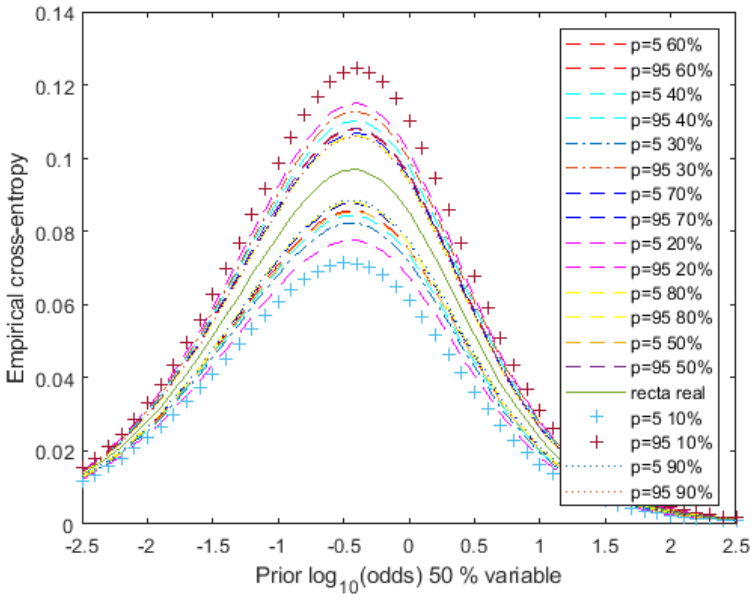

Figura A.6: *Curva entropía óptima y sus correspondientes intervalos de confianza del 90 % de acierto con bootstrap para 1000 repeticiones y cogiendo desde el 90 % al 10 % de 10 en 10 de la muestra inicial.*

B

Anexo B: Subset Bootstrap

B.1. Ejemplo subset bootstrap de todos los valores de una curva

En este apartado se muestran los resultados obtenidos variando la cantidad de la muestra subset bootstrap. En la siguiente figura mostraremos los intervalos de confianza cogiendo desde el 10 % de los datos hasta el 90 % de la muestra modificada y veremos como a medida que aumenta este porcentaje se va reduciendo el intervalo:

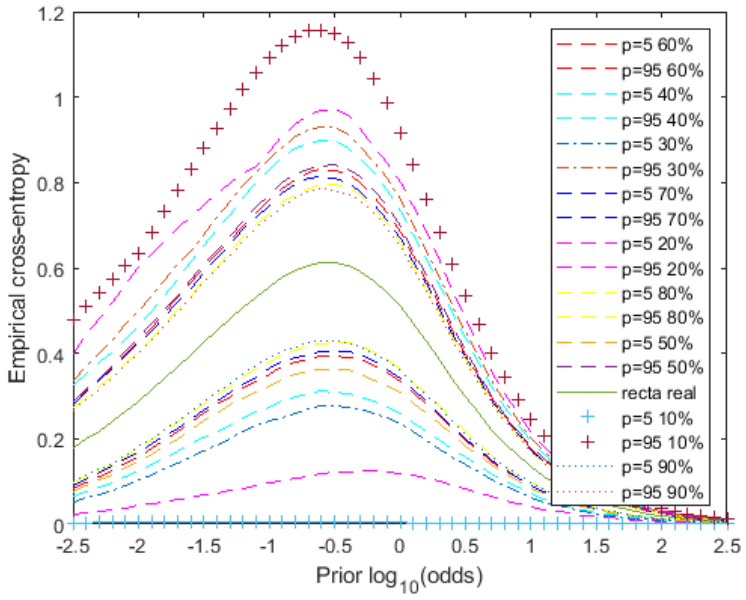

Figura B.1: *Curva entropía cruzada y sus correspondientes intervalos de confianza del 90 % de acierto con subset bootstrap para 1000 repeticiones y cogiendo desde el 90 % al 10 % de 10 en 10 de la muestra inicial.*

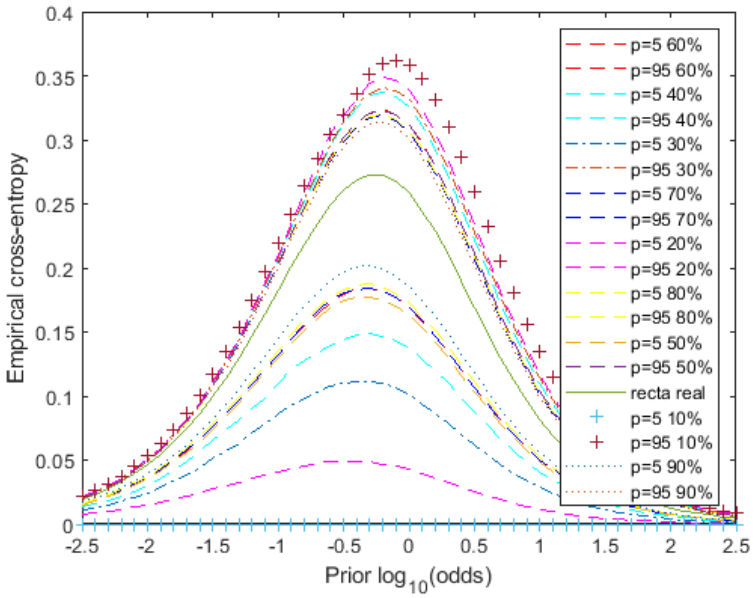

Figura B.2: *Curva entropía óptima y sus correspondientes intervalos de confianza del 90 % de acierto con subset bootstrap para 1000 repeticiones y cogiendo desde el 90 % al 10 % de 10 en 10 de la muestra inicial.*